PREDICTING WATER POTABILITY

Naveen Parthasarathy

Table of Contents

Introduction

- The lack of access to safe and clean water is one of the largest risk factors for the spread polio, typhoid, cholera etc.
- The unavailability of drinkable water worsens and intensifies malnutrition, particularly in children.
- Treatment to make water potable is also not cheap, and in most cases, affect a large number of people.

Problem Statement

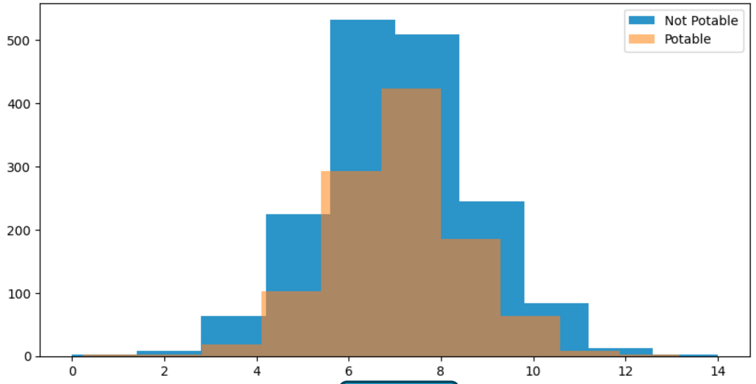
- In this project, we will explore the use of machine learning algorithms to predict the potability of water based on various physical and chemical properties.
- Our goal is to build a machine learning model that can accurately predict the potability of water based on these properties.
- This model can be used to provide preliminary information on the potability of a water sample.

Methodology – Data Sourcing

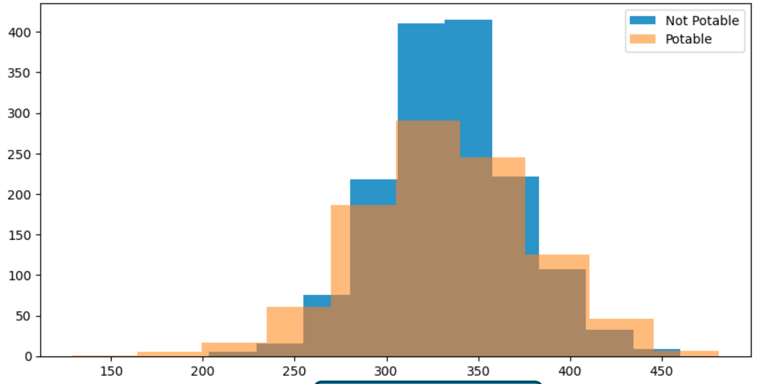
- The dataset obtained was from Kaggle. Link
- It consists of 3276 water samples, including information such as pH, conductivity, and other physical and chemical properties.

Methodology – Data Exploration

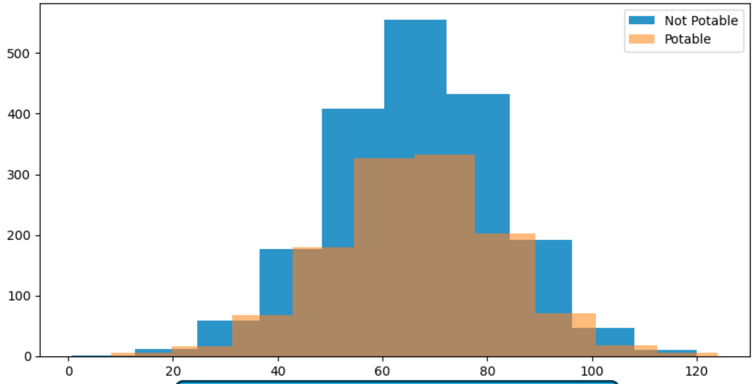
- All datatypes were of type float. The target is called Potability and is a binary variable.
- 5 Columns – Hardness, Solids, Chloramines, Conductivity, Organic_carbon and Turbidity had no missing values.



ph



Sulfate



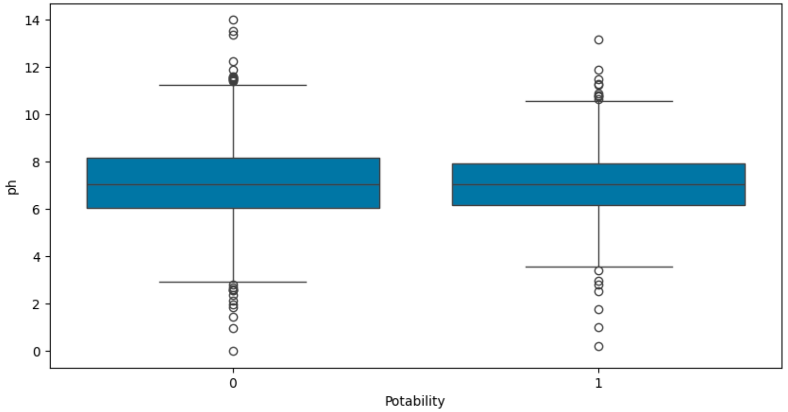
Trihalomethanes

- All columns were normally distributed, except Solids.

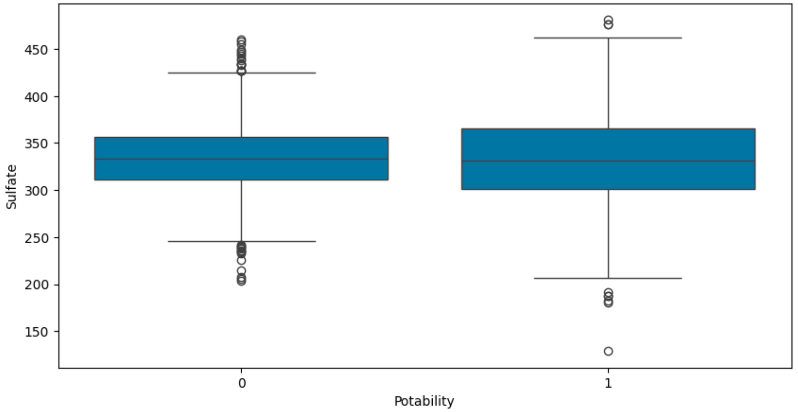
Methodology – Data Exploration

- All columns possessed outliers when stratified into Potable and Non-Potable samples.

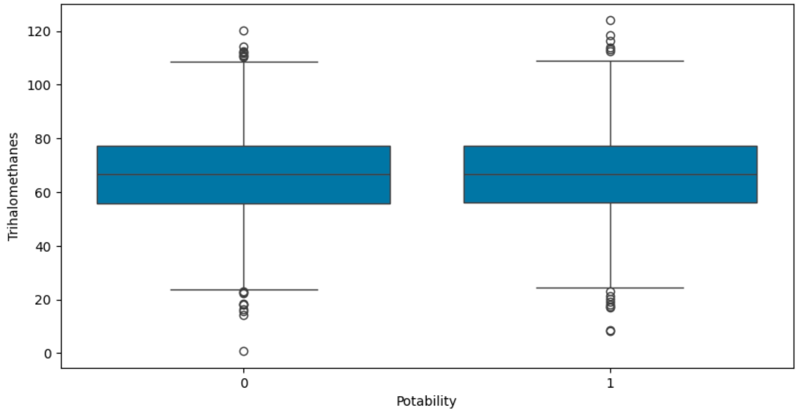
ph



Sulfate

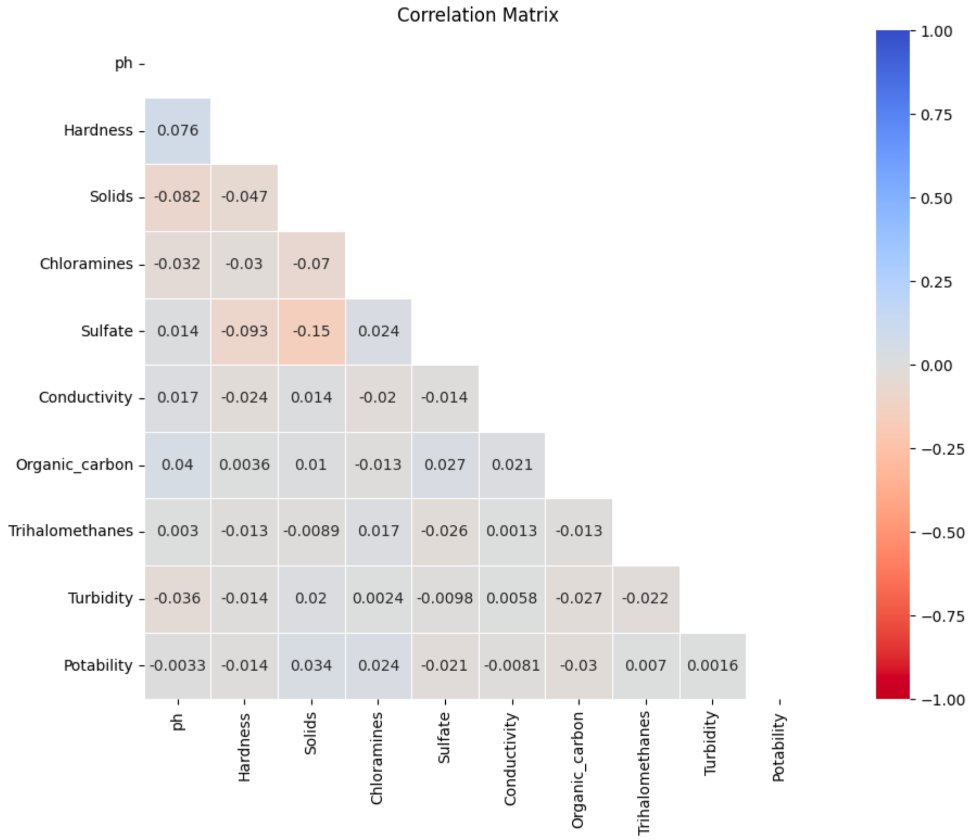


Trihalomethanes



Methodology – Data Exploration

- Columns were all converted to the same unit of measurement (ppm).
- Missing data was imputed with a KNN-imputer
- A correlation matrix was created to check for linear correlation between variables.

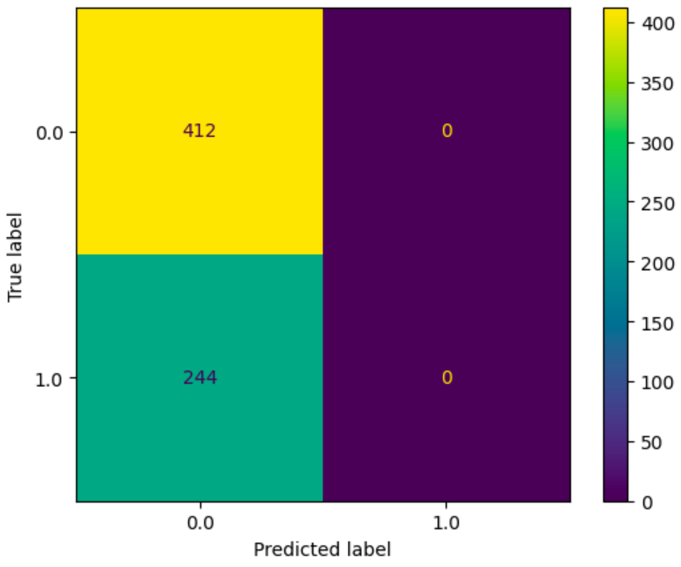


Methodology - Models

- The metrics to focus on are accuracy and recall.
    - Accuracy since it is important to know out of all predictions, which water samples are actually potable.
    - Recall since it is important to know out of all potable water samples, which ones are predicted to be potable.
- Based on the EDA so far, Linear models may not be useful to determine potability.
- The use of more effective classifiers, like Random Forest and Neural Network may lead to better accuracy.
- The data was split into training and testing sets, with the test set being 25% of the size of the data.

Methodology - Models

- Baseline
  - Simple model that checked the proportion of potability (1) vs. non-potability (0).
  - 61% of samples were not-potable.
  - Therefore, predicting 0 would yield a 61% accuracy.
  - Goal: Create a model that > 61% accuracy
- Logistic Regression
  - Assumption: This model performs poorly
  - Pipeline consisted of a standard scaler, normalizer and logistic regression model.
  - The normalizer is used to handle outliers.
  - Grid searched the ‘C’ parameter.
  - Accuracy: 62.8%, Recall: 100% on non-potability – 0% on potability.



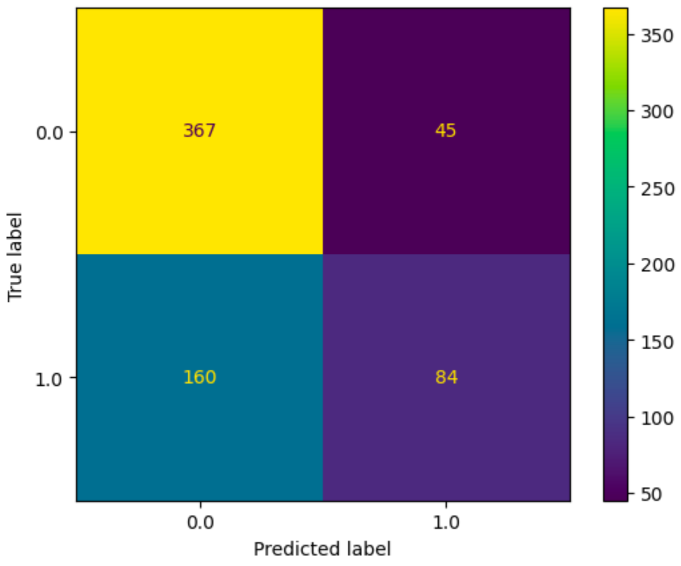
Only predicts 0, just like baseline model

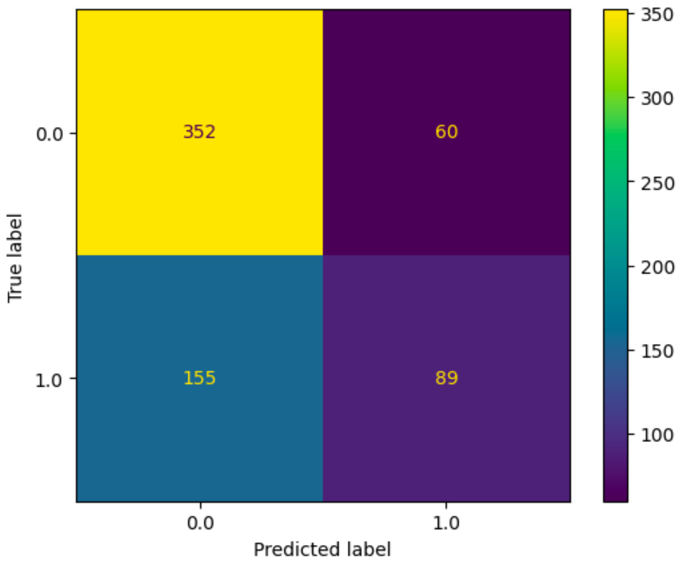
Methodology - Models

- Random Forest Classifier
  - Similar pipeline to the logistic regression.
  - Grid searched parameters are ‘n_estimators’ and ’max_depth’.
  - Accuracy: 68.75%, Recall: 89% on non-potability – 34% on potability.

- Simple Neural Network
  - Data is only scaled.
  - Consists of 1 hidden layer.
  - Accuracy: 66%, Recall: 40% on potability – 81% on non-potability.

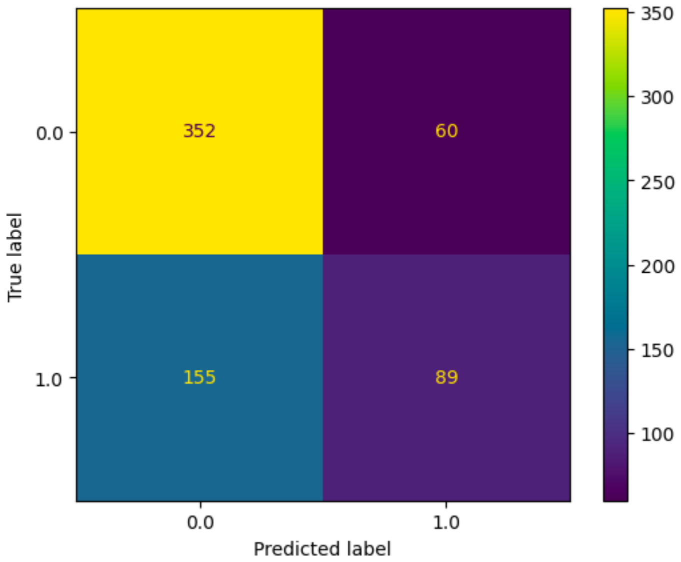
RF confusion matrix





Simple NN confusion matrix

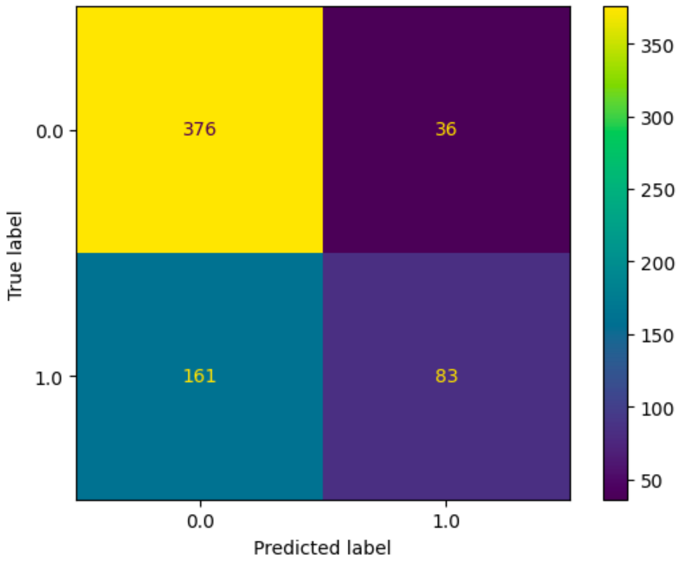
GSNN confusion matrix



Methodology - Models

- Grid Searched Neural Network
    - Data is scaled and normalized through pipeline.
    - Consists of a dropout layer as well as a hidden layer.
    - Grid search on ‘batch_size’, ‘epochs’ and ‘optimizer’.
    - Accuracy: 67%, Recall: 36% on potability, 85% on non-potability.

- RF with polynomial features
    - Similar pipeline with the addition of polynomial features.
    - Grid search includes ‘interaction_only’, ‘include_bias’ and ‘degree’.
    - Accuracy: 70%, Recall: 91% on non-potability, 34% on potability.



RF with poly feats

Inferences

- Range of models from 61% accurate to 70% accurate.
- Non-linear relationship between potability and water quality features.
- Live demonstration of project available here - Link

Future Work

- Feature Engineering
  - Create/dummify features from the dataset.
- Models like LGBM, XGBoost etc.
- Data Augmentation
  - Synthetic data using SMOTE or adding more water samples.
- Hyperparameter tuning
  - Large number of parameters can be tuned to find optimal model

References

- www.Kaggle.com
- www.streamlit.io

THANK YOU