Documentation for Water Quality Machine Learning Project

Background:

A water treatment plant in Mountain View (CA) has been serving water that did not meet the requires stipulated by the Safe Water Drinking Act in America. The penalties associated with the violation is declared in the website:

https://www.epa.gov/enforcement/safe-drinking-water-act-sdwa-and-federal-facilities

Excepts:

The 1996 SDWA amendments clearly express EPA 's administrative penalty authority over federal agencies in section 1447, 42 U.S.C. §300j-6. SDWA section 1447 provides the blanket authority for EPA to issue penalty orders to federal agencies for violations of any requirement of the SDWA or a requirement or schedule imposed by an administrative compliance order, an imminent and substantial endangerment order, or other administrative order issued under the SDWA.
SDWA penalties against federal agencies may be up to $32,500 per day per violation after January 12, 2009. (The Federal Civil Penalties Inflation Adjustment Act of 1990 requires EPA to revise every four years the penalty amounts available under federal environmental statutes, including the SDWA. For the current penalty amounts, see 40 C.F.R. Part 19, Table 1 of Section 19.4 .

The stakeholders are looking for a solution to deal with the problem.

Data availability:

The stakeholders provided around 3000 samples of data consisting of 9 water properties and each comes with an outcome of either '0' or '1'. '0' being undrinkable and '1' being drinkable.

Methodology:

Raw data provided by stakeholders will be pre-processed. This includes fixing imbalanced data between undrinkable water (class '0') and drinkable water (class '1') and standardizing the dataset. After pre-processing, we will use 2 methods to arrive at the top 3 classifiers. The ensemble method and the optimized pipeline method. The ROC AUC score will be used to determine the top 3 classifiers.

Ensemble Method:
We will combine 8 different classifiers to arrive at one main classifier (ie. StackingClassifier). The purpose is to arrive at a classifier which will give the highest ROC AUC score. This score indicates the probability of the classifier in separating undrinkable water from drinkable water.
These 8 classifiers are RandomForest, Support Vector Machine, DecisionTree, RidgeClassifier, KNearest, NaiveBayes, Multi_LayerPerceptron (aka neural networks) and XGBoost. StackingClassifier will be used to consolidate all these classifiers to arrive at one classifier aka StackingClassifier using LogisticRegression as the final estimator (aka Meta Learner).

Prepared by: Chan Chong Meng | Email: chongmeng.chan@gmail.com

Optimized Pipeline:
At same time, we will use an automated method aka Optimized Pipeline. This will give search through all available classifiers to arrive at the best combination of these classifiers. This Optimized Pipeline is known as TPOTClassifier. We will apply the same score (ROC AUC score) to this classifier.

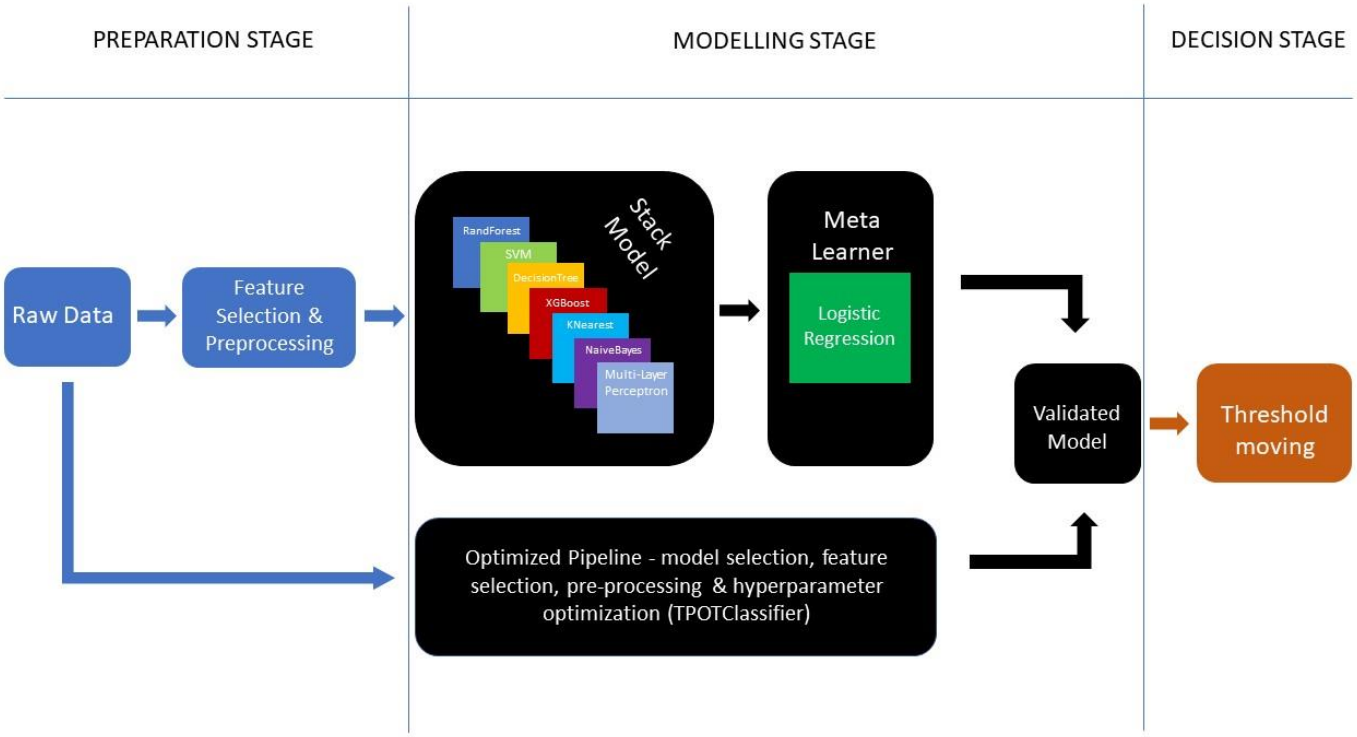
At this point we will have 8 classifiers, the StackClassifier and the TPOTClassifier. In total there will be 10 classifiers. We will rank all these classifiers according to their ROC AUC score and select the top 3 classifiers.

Threshold Moving with Specificity score as the key determinant:
After we have arrived at the top 3 classifiers, this will come to decision stage. Here the stakeholders will decide which thresholds to use to determine specificity.
In other words, the stakeholders will apply Threshold Moving to arrive at a classifier which will meet their requirements. Threshold ranges between 0.2 to 0.9.
Moving the thresholds higher will result in higher benchmark being used for identifying drinkable water, thus resulting in more undrinkable water being identified. This will eventually lead to minimising the cost of misclassification (ie. Misclassifying undrinkable water as drinkable aka 'False Positive').
The purpose of moving the thresholds is to identify undrinkable water (ie. class '0'), aka **'Specificity'**.
Specificity will change as thresholds move. Therefore, the stakeholders will determine the threshold which will give them a comfortable Specificity to minimize the risk/cost of misclassification.

In summary, we will utilize the **Ensemble method (StackingClassifier)** as well as the **Optimized Pipeline method (TPOTClassifier)** to arrive at the top 3 models (based on ROC AUC score) with the best capabilities in separating drinkable water from undrinkable water.

Prepared by: Chan Chong Meng | Email: chongmeng.chan@gmail.com

Methodology Diagram:



Key processes and deliverables by machine learning specialist:

1. Raw data provided by stakeholders will be pre-processed before implementing the ensemble method (Stacking Classification).
2. At the same time, the raw data will also be used for Optimized Pipeline method (TPOTClassifier).
3. ROC AUC readings will be taken, where the top 3 classifiers will be selected to go for Threshold Moving in the decision stage.
4. Metrics used will be specificity, and sensitivity, of which special emphasis will be given to specificity. These will be measured across all thresholds (0.2 – 0.9).
5. A specificity / sensitivity curve plot against all thresholds (from 0.2 to 0.9) will be provided to the stakeholders as this will give them better understanding through visualization for them to make a best decision on thresholds.
6. To immediately output the batch numbers for drinkable waters (ie. Class '1') and undrinkable waters (ie. Class '0') based on desired thresholds.

Conclusion:

The top 3 models are StackingClassifier, TPOTClassifier and XGBoost based on ROC AUC readings. Our deliverables end here.

Threshold determination will be the decision of the stakeholders.

Prepared by: Chan Chong Meng | Email: chongmeng.chan@gmail.com