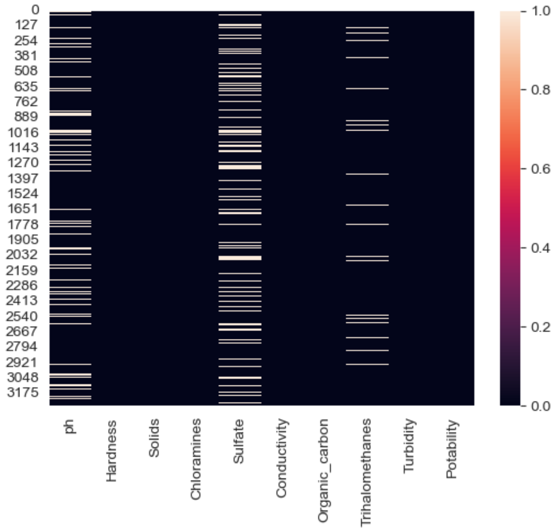
ML coursework Supplementary material

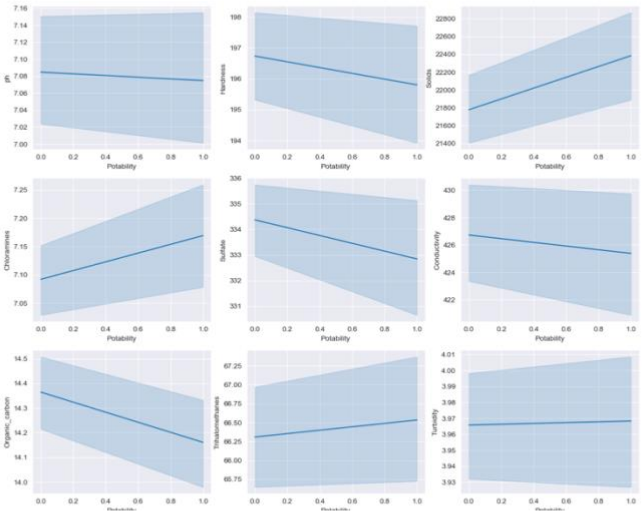
This PDF file shall mention the exploratory analysis and finding that is not worth including in the poster, some of the analysis provides interesting insight leading us to the final poster and final construct of our model and also play an important part in this project to carry on in the right direction.

Handle with missing value & Graph of missing data frequency



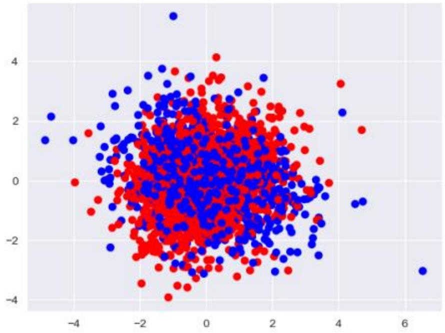
From the figure above some of the variables contain missing values we tried many ways to handle these missing values, and in the end, we choose the method that imputes the mean in missing value rows as we thought the most appropriate way. The reason is filling the missing values with 0 value can decrease model accuracy due to water quality being sensitive data we should not impute 0 in the missing row and removing missing value rows can lead to the results in much less data on the model training set, resulting in inaccurate models.

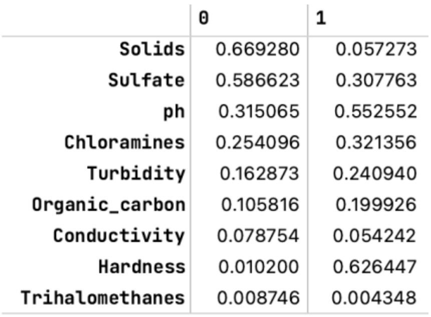
The correlation between water features and target value



We represent the correlation in the line graph to see the relation between other features and the target value in this project our target value is an ordinal number 1 and 0 that we think presenting it like a boxplot is a better way to represent this value and compare the relation between target classes to see which features importance to select for training our model.

Principal component analysis (PCA)





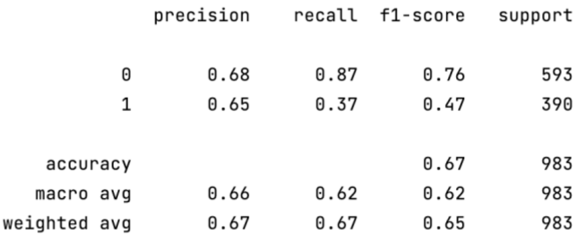
	0	1
Solids	0.669280	0.057273
Sulfate	0.586623	0.307763
ph	0.315065	0.552552
Chloramines	0.254096	0.321356
Turbidity	0.162873	0.240940
Organic_carbon	0.105816	0.199926
Conductivity	0.078754	0.054242
Hardness	0.010200	0.626447
Trihalomethanes	0.008746	0.004348

PCA is another method to study and understand the patterns of the data except correlation heatmap and correlation graph for reducing the dimensionality of a dataset into two components to see the trend of our data and which feature has the most impact on this dataset. The first two components capture our data only 25% and we impute color to represent the difference of the prediction classes (red: undrinkable water, blue: drinkable water) which we cannot tell anything so much about this. However, we can bear in mind that for the first component solid and sulfate variables quite influence this dataset. On the other hand, hardness and trihalomethanes are inconsistent with this dataset.

intermediate results & other model construction

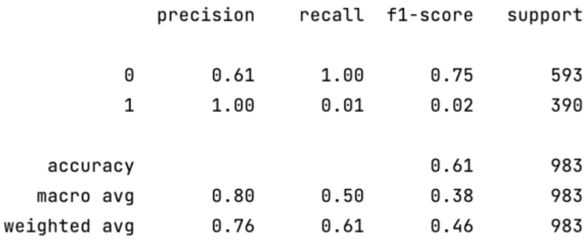
Unsurprisingly, the Random Forest model shows the result of the highest accuracy when predicting the potability of water from this dataset. However, Naïve Bayes and logistic regression also provide an acceptable result when making a prediction. To carry on our analysis plan in the end we choose random forest and decision tree as two models we have to compare.

Naïve bayes



```
              precision    recall  f1-score   support

           0       0.68      0.87      0.76       593
           1       0.65      0.37      0.47       390

    accuracy                           0.67       983
   macro avg       0.66      0.62      0.62       983
weighted avg       0.67      0.67      0.65       983
```

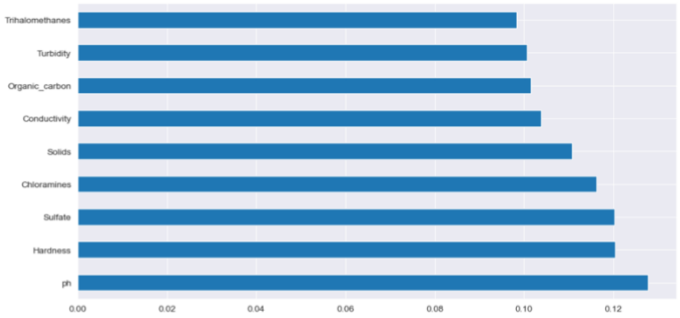
Logistic regression

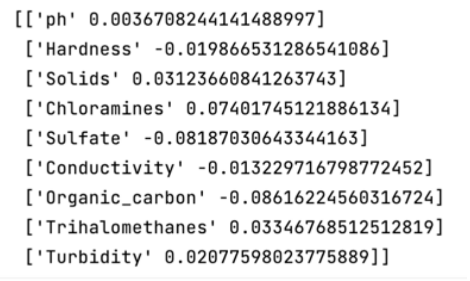


```
              precision    recall  f1-score   support

           0       0.61      1.00      0.75       593
           1       1.00      0.01      0.02       390

    accuracy                           0.61       983
   macro avg       0.80      0.50      0.38       983
weighted avg       0.76      0.61      0.46       983
```

For other model results above, we produce different models to compare with our main model (random forest and decision tree) performance and to find the important feature to reduce the feature before building the comparison model.

Feature importance random forest model & coefficient score Logistic regression



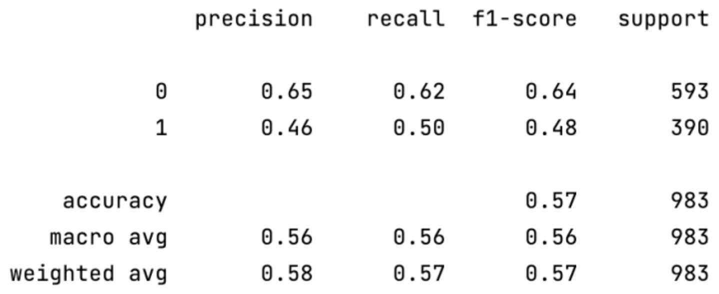


```
[['ph' 0.0036708244141488997]
 ['Hardness' -0.019866531286541086]
 ['Solids' 0.03123660841263743]
 ['Chloramines' 0.07401745121886134]
 ['Sulfate' -0.08187030643344163]
 ['Conductivity' -0.013229716798772452]
 ['Organic_carbon' -0.08616224560316724]
 ['Trihalomethanes' 0.03346768512512819]
 ['Turbidity' 0.02077598023775889]]
```

Logistic regression model does not perform well to predict class 1 so from the figure above we should not conclude from the coefficient score (right picture) for the feature selection.

Besides, the interesting thing that we know from this section is random forest model weight our feature quite equal to the initial building model. Therefore, in the end, we decided to put all water quality features into our final model.

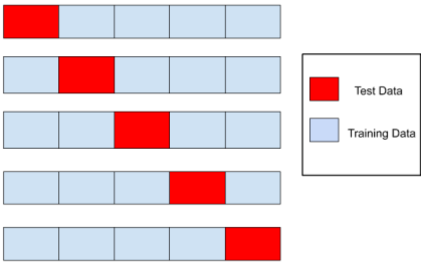
Decision tree performance before tuning hyperparameter



```
              precision    recall  f1-score   support

           0       0.65      0.62      0.64       593
           1       0.46      0.50      0.48       390

    accuracy                           0.57       983
   macro avg       0.56      0.56      0.56       983
weighted avg       0.58      0.57      0.57       983
```

As we can see decision trees produce fewer accuracy scores compared to other models but in order to follow the plan that we had set out to compare RF and DT we need to apply the tuning hyperparameter technique to enhance the model to produce the best model for comparison with Random Forest model. However, the process of constructing the model and improving we shall carry on in Matlab and mentioned in the poster.

k-fold cross validation (we used 10-fold in this project as a default)



As for the poster, we didn't explain this part as deeply as it should we assume that this part is a piece of basic knowledge, and the main priority is to compare between two models.

Using n-fold CV results in different models for each set of model hyperparameters. Once I have worked out which set of model hyperparameters minimize the kfold loss score we selected that model for building our final model.

It is worth mentioning in the supplementary that we applied k-fold cross-validation to validate our model and prevent our model from working well only training data lead to poor performance in the unseen data. Therefore, k-fold cv had been applied in this project to validate and check the performance of the model when tuning hyperparameters to produce the final model for comparison in the end.