Water Potability

Shen Chenyu:1155206444 Cui Changrui: 1155201565 Mao Xuren: 1155208112
Liu Yaorui: 1155206042 Xiang Yue: 1155205208 Zhang Jing: 1155207951

1. Introduction and EDA

Due to human activities, the previously scarce potable water on Earth is being increasingly polluted. Although Hong Kong is surrounded by the sea on three sides, there is no major river within its territory, so freshwater resources are very scarce. Potable water is associated with our everyday life and the development of the society. How to reduce water pollution has become a hot topic of social concern.

(1) NA Replacement[1]

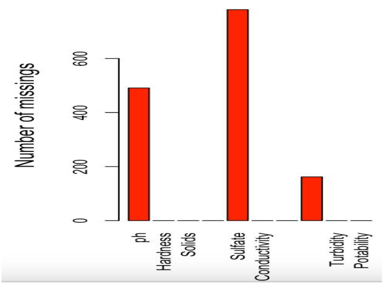


figure1.1

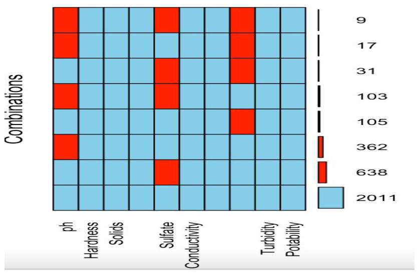


figure1.2

From the picture, we can find that most of the NA values are distributing in “ph” column and “Sulfate” column. Considering the NA values are so many , we take use of the mean value of each attribute to replace the NA values.

(2) Outliers

We use the boxplot to detect the outliers. Values that are out of the boundray of 1.5*IQR are treated as the outliers.

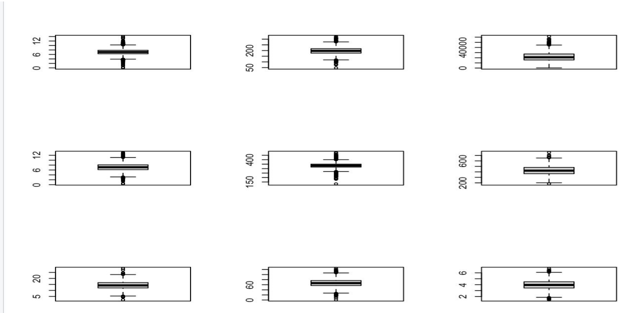
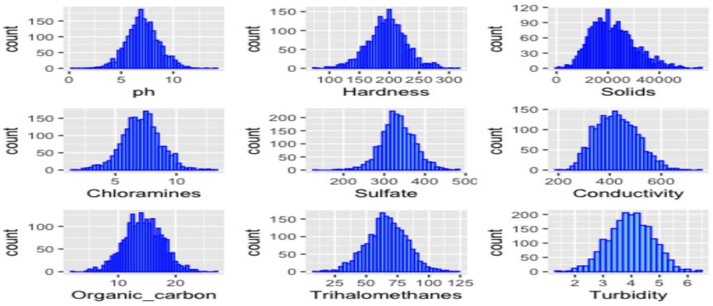


figure1.3

(3) Descriptive analysis of independent variables



[1] Datasourses:https://www.kaggle.com/datasets/uom190346a/water-quality-and-potability

Since all the independent variables are continuous, we use histogram to show the distribution of these variables. Most of them could be considered as approximately normal distribution.

2. Variable filtering

Because there are too many attributes in the dataset, we want to select some attributes which have greater degree of influence on potability to be the independent variable. There are various of method to filter variables, such as LASSO, PCA, Information Value and so on. In this case, if the factors are largely correlated, we will use principal component analysis(PCA). Otherwise, we will calculate the Information Value of each factor to select the most three significant factors.

figure1.4

(1) Relevance Testing

We use R to do the relevance testing, the results of the outputs are shown in the figure below.

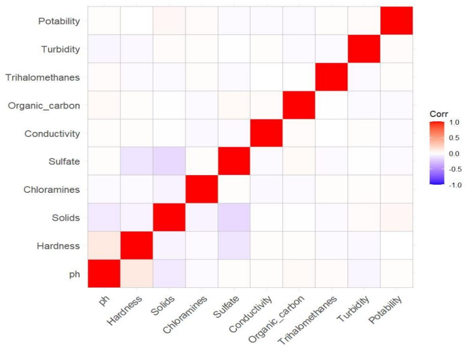


figure2.1

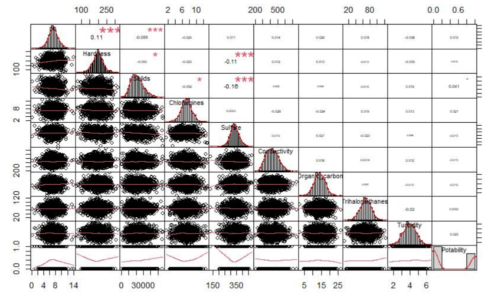
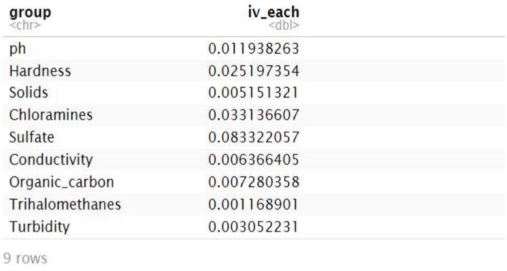


figure2.2

In the figure 2.1, orange means positive correlation, purple means negative correlation, and darker color means higher correlation. We can see that even the darkest colored area is below 0.5. More accurately, we can look at Figure 2.2, and the value between the two variables with the highest correlation is 0.16, which is too low. So, to a certain degree, the variables are independent to each other.

(2) Information Value Calculating

We use R to do the calculating of Information Value, and the results are in the figures below.



group <chr>	iv_each <dbl>
ph	0.011938263
Hardness	0.025197354
Solids	0.005151321
Chloramines	0.033136607
Sulfate	0.083322057
Conductivity	0.006366405
Organic_carbon	0.007280358
Trihalomethanes	0.001168901
Turbidity	0.003052231

9 rows

table2.1

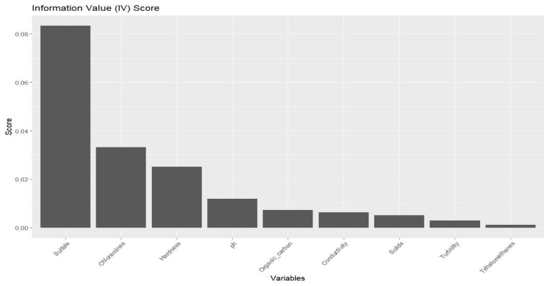
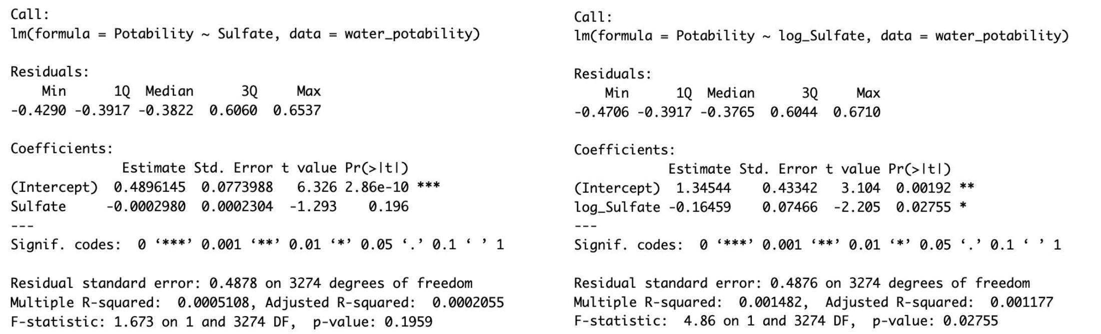


figure2.3

According to the Table 2.1 and Figure 2.3, we can see that the most three significant factors are sulfate, Chloramines and Hardness.

3. Linear Regression and Bayes

The linear regression of potability on sulfate has an intercept of 0.48 and a slope of −0.0002, indicating a negative correlation. The R-squared of regression is 0.0002. This means that 0.02% of the overall variation in potability is explained by the linear regression with Sulfate, while remaining 99.98% is left unexplained, which is very low, and the adjusted R-square is negative, suggesting poor model fit. Even with a level-log transformation, the resulting model tends to have a low R-value, indicating a minimal improvement in explaining the variability in Potability. The linear regression can predict values outside the range of 0 and 1, given that our dependent variable is binary, linear regression may not exhibit discriminative patterns. Logistic regression is commonly recommended, which will fit a curve that represents the probability of potability. However, some references suggest considering a Bayesian approach as an alternative.



```
Call:
lm(formula = Potability ~ Sulfate, data = water_potability)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4290 -0.3917 -0.3822  0.6060  0.6537 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.4896145  0.0773988   6.326 2.86e-10 ***
Sulfate     -0.0002980  0.0002304  -1.293    0.196    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4878 on 3274 degrees of freedom
Multiple R-squared:  0.0005108, Adjusted R-squared:  0.0002055 
F-statistic: 1.673 on 1 and 3274 DF,  p-value: 0.1959
```

```
Call:
lm(formula = Potability ~ log_Sulfate, data = water_potability)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4706 -0.3917 -0.3765  0.6044  0.6710 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  1.34544    0.43342   3.104  0.00192 **
log_Sulfate -0.16459    0.07466  -2.205  0.02755 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4876 on 3274 degrees of freedom
Multiple R-squared:  0.001482,  Adjusted R-squared:  0.001177 
F-statistic:  4.86 on 1 and 3274 DF,  p-value: 0.02755
```

Figure 3.1 Potability and Sulfate & log—Sulfate: Linear regression & level—log Linear Regression

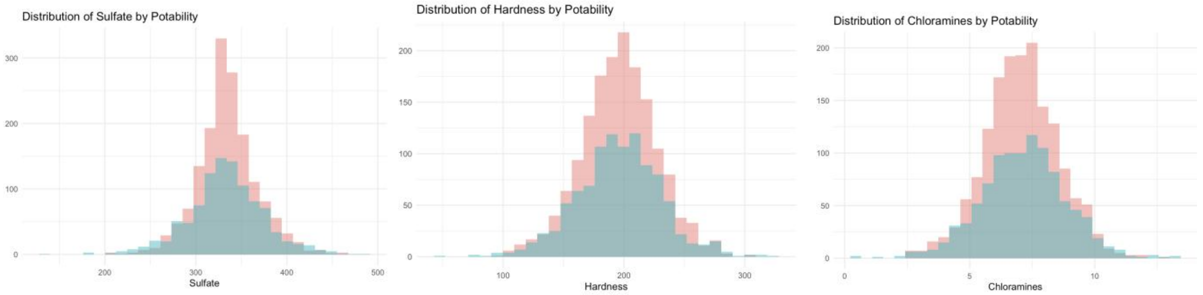
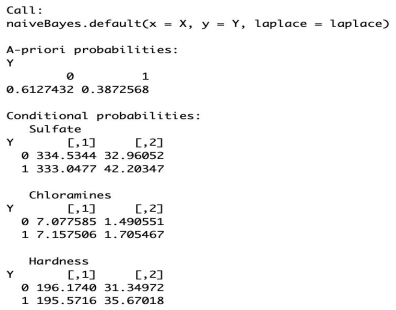


Figure 3.2 Distribution of Sulfate, Hardness and Chloramines to Potability



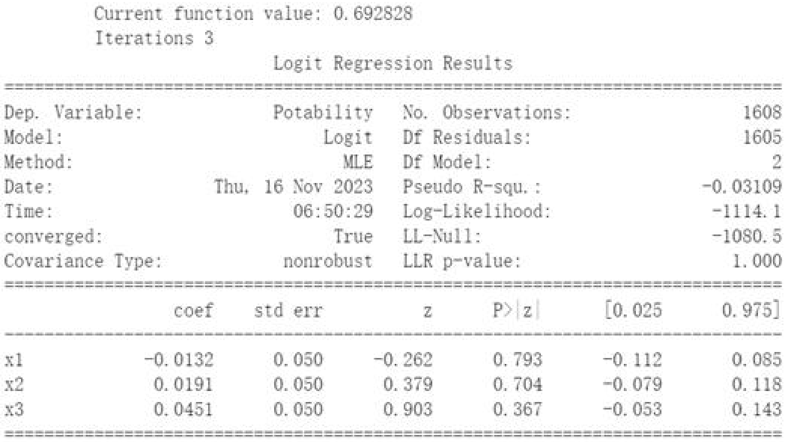
```
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1 
0.6127432 0.3872568 

Conditional probabilities:
   Sulfate
Y       [,1]     [,2]
  0 334.5344 32.96052
  1 333.0477 42.20347

   Chloramines
Y       [,1]     [,2]
  0 7.077585 1.490551
  1 7.157506 1.705467

   Hardness
Y       [,1]     [,2]
  0 196.1740 31.34972
  1 195.5716 35.67018
```

Figure 3.2 Distribution of Sulfate, Hardness and Chloramines to Potability

In the Naive Bayesian analysis, we found a 61% chance for observations to be classified as non-potable (class 0) and a 38.7% chance for potable (class 1). It is clear that these mean and standard deviation values are quite similar within two classes. Graphical representations reinforce this, indicating significant overlap in the distributions, towards Gaussian distributions. As a result, these features seem less informative for the model. It's important to note that Naive Bayes doesn't calculate the Bayesian Information Criterion (BIC) since it operates differently from linear models and doesn't directly use likelihood functions.

4. Logistic Regression

(1) Logistic Regression Results



```
         Current function value: 0.692828
         Iterations 3
                           Logit Regression Results
==============================================================================
Dep. Variable:             Potability   No. Observations:                 1608
Model:                          Logit   Df Residuals:                     1605
Method:                           MLE   Df Model:                            2
Date:                Thu, 16 Nov 2023   Pseudo R-squ.:                -0.03109
Time:                        06:50:29   Log-Likelihood:                -1114.1
converged:                       True   LL-Null:                       -1080.5
Covariance Type:            nonrobust   LLR p-value:                     1.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0132      0.050     -0.262      0.793      -0.112       0.085
x2             0.0191      0.050      0.379      0.704      -0.079       0.118
x3             0.0451      0.050      0.903      0.367      -0.053       0.143
==============================================================================
```

figure4.1

(a) P>|z |: None of the variables (Sulfate, Hardness, Chloramines) have p-values less than 0.05. This suggests that these variables may not have a significant impact on determining the potability of water.

(b) Coefficient: In this case, the coefficient of x1 is -0.0132, the coefficient of x2 is 0.0191, and the coefficient of x3 is 0.0451. The values of the coefficients are all relatively close to zero, which suggests that the relationship between the independent variable and the dependent variable is weak.

(c) std error: In this case, the standard errors for all three independent variables are 0.050, indicating relatively low uncertainty in the coefficient estimates.

(2) Confusion Matrix

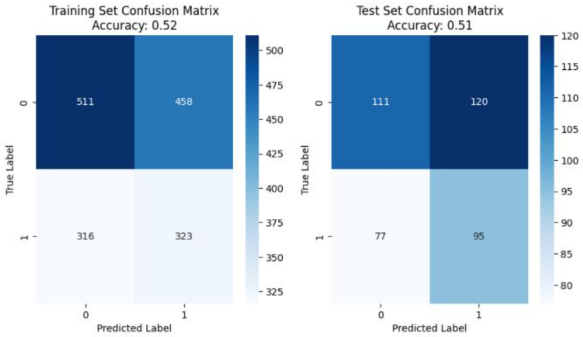


figure4.2

To evaluate the performance of the model, we can look at the accuracy. On the training set, the accuracy is 0.52, indicating that our model correctly classified 52% of the samples. On the test set, the accuracy is 0.50, which means that the model's predictive power is close to random guessing.

(3) Other Performance Evaluation Metrics

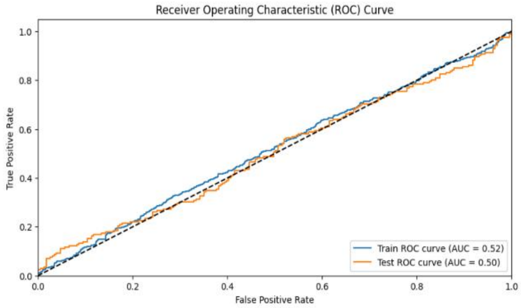


figure4.3

Both the training set AUC and the test set AUC are presented. The training set AUC of 0.52 suggests that the model has some discriminatory power in the training data, while the test set AUC of 0.50 indicates that the model's discriminatory power is close to random guessing on unseen data. Further analysis and improvement may be necessary to enhance the model's performance.

5. Random Forest

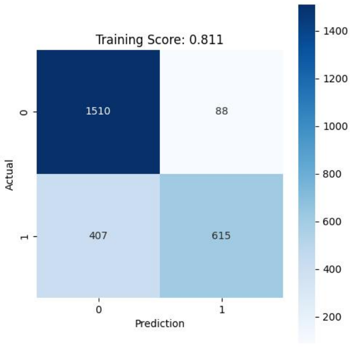


figure5.1

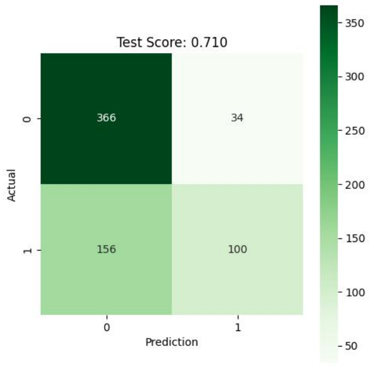


figure5.2

This has resulted in a certain level of false positives, as shown in the confusion matrix, with 407 potable samples wrongly classified as non-potable. The confusion matrix further corroborates this, with 1510 non-potable samples and 615 potable samples correctly predicted. These figures provide us with a baseline for the model's performance during training and set expectations for its performance on the test set.

Turn our attention to the test set. On the test set, the model maintains a relatively high precision for predicting non-potable water, with a recall of 0.92, meaning 366 out of 400 non-potable water samples are correctly predicted. However, for potable water, the model's performance declines. Out of 256 potable water samples, only 100 are correctly predicted, indicating a considerable number of potable samples being misclassified as non-potable.

Overall, the model has an accuracy of 0.71 on the test set, a decrease from the training set's 0.81, reflecting a certain loss of performance on unknown data. Despite the highly accurate predictions for figure 5.1, the significant number of false positives (156) for figure 5.2 suggests we need to find a better balance between predictive power and false positives.

Summarizing the test set performance, we can say that while the model excels at identifying non-potable water, there is room for improvement in confirming the safety of potable water.

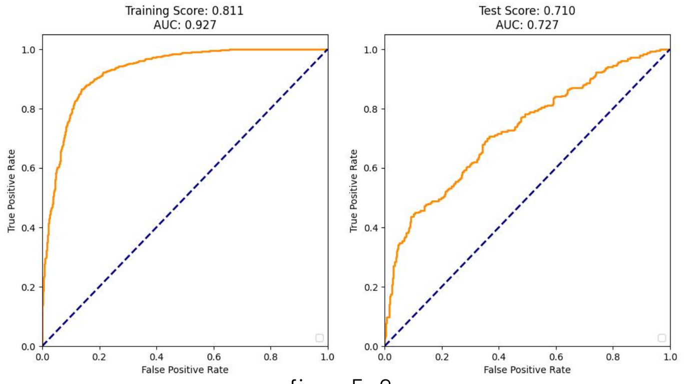


figure5.3

Our Random Forest model showed excellent performance on the training set with an AUC value of 0.927, indicating a very high accuracy in distinguishing between non-potable and potable water samples. Most of ROC curve lies above the diagonal line, indicating good discriminatory ability. However, the AUC value of test set drops to 0.727. The test set's ROC curve, closer to the diagonal line compared to the training set, confirms this performance decrease on unseen data.

6. SVM

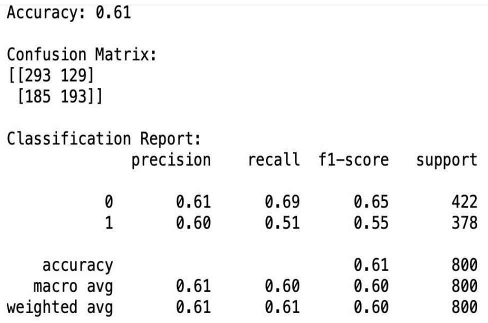
Considering the issue of Random Forest overfitting, with an AUC of 0.927, SVM proves apt for binary classification. Our research focuses on water probability, a binary query.

The SVM steps involve:

(a) Initially, feature standardization ensures uniform scaling across all features.

(b) The model is trained using a Radial Basis Function kernel, apt for non-linear challenges, as SVM typically employs this kernel.

(c) Predictions are then made on the test set. The final output is shown in Figure 6.1.



```
Accuracy: 0.61

Confusion Matrix:
[[293 129]
 [185 193]]

Classification Report:
              precision    recall  f1-score   support

           0       0.61      0.69      0.65       422
           1       0.60      0.51      0.55       378

    accuracy                           0.61       800
   macro avg       0.61      0.60      0.60       800
weighted avg       0.61      0.61      0.60       800
```

figure6.1

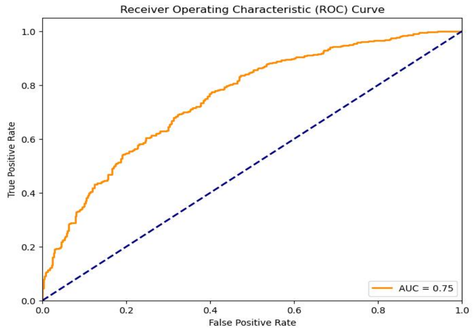


figure6.2

With a precision of 0.61 for the positive class, this implies that 61% of predictions are indeed positive. The AUC value of 0.75 in SVM suggests commendable model performance in predicting water probability in binary classification.

7. Model evaluation and conclusion

Model	recall	f1-score	precision	AUC
Naive bayes	0.50	0.5	0.62	0.50
Logistic regression	0.40	0.45	0.49	0.52
Random Forest	0.77	0.79	0.82	0.927
SVM	0.67	0.67	0.67	0.75

figure 7.1

I In summary, the Random Forest model (AUC 0.927) outperforms others, with SVM (AUC 0.75) also showing robust results. Conversely, Naive Bayes (AUC 0.50) and Logistic Regression (AUC 0.52) exhibit comparatively weaker performance.

II Further consideration： water quality significantly varies based on geographical location and sources. Sole reliance on binary labels (potable or non-potable) for water quality assessment is overly simplistic. This necessitates more nuanced segmentation.

III Model optimization a recommendations

(a) Random Forest Regression: Adjust the forest's maximum depths to 5, 8, 10, and the total number of trees to 200, 400.

(b)Employ unsupervised Learning, like K-means, ideal for exploratory data analysis in novel domains. Given the diverse water quality, better segmentation than binary labels is required.

IV Conclusion

In summary, key lessons include the criticality of meticulous data cleaning and preprocessing, comprehending assumptions of various models, and the imperative of continual enhancement and optimization in machine learning. This entails potentially integrating diverse algorithms to identify the optimal approach for tasks like classification, regression, or other types of problems.