Project_Restivo_Drury_Cirnigliaro

Francesco Restivo, Thomas Alfio Drury, Fulvio Umberto Cirnigliaro

Preliminary Questions

1. Is the dataset artificially generated?

2. Which model performs best in terms of accuracy?

3. Which model is best for predicting drinkable water?

Introduction

The dataset contains information on the potability of water, helping to identify key factors that determine whether water is safe for consumption. It consists of n = 1608 observations and p = 9 variables; the response variable is "Potability".

The variables are described below.

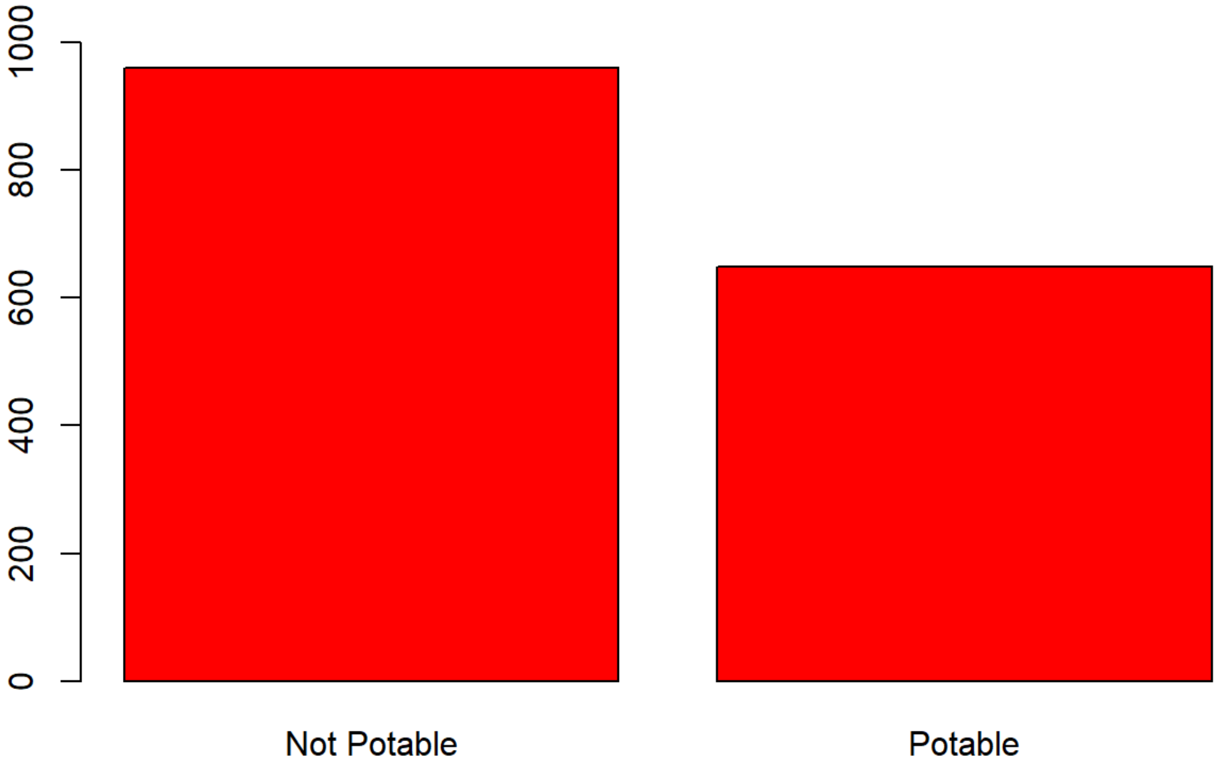
Variables	Type	Description	Measurement unit
PH	Numeric	A parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status.	mol/Kg
Hardness	Numeric	Expresses the salts that are dissolved from geologic deposits through which water travels. Mainly caused by calcium and magnesium.	mg/L
Solids	Numeric	Minerals, dissolved in water, that produce unwanted taste and diluted color in appearance of water.	ppm
Chloramines	Numeric	Chloramines are formed when ammonia is added to chlorine to treat drinking water.	ppm
Sulfate	Numeric	Sulfates are naturally occurring substances found in minerals, soil, and rocks, used mainly in	mg/L

		the chemical industry.	
Conductivity	Numeric	Increase in ion concentration enhances the electrical conductivity of water.	µS/cm
Organic Carbon	Numeric	TOC measures the total amount of carbon in organic compounds in pure water, typically from decaying natural organic matter.	ppm
Trihalomethanes	Numeric	THMs are chemicals found in chlorinated water, affected by organic content, chlorine level, and temperature.	µg/L
Turbidity	Numeric	Turbidity depends on the amount of solid matter suspended in water.	NTU
Potability	Factor	Indicates if water is safe for human consumption: 1 = Potable, 0 = Not potable.	

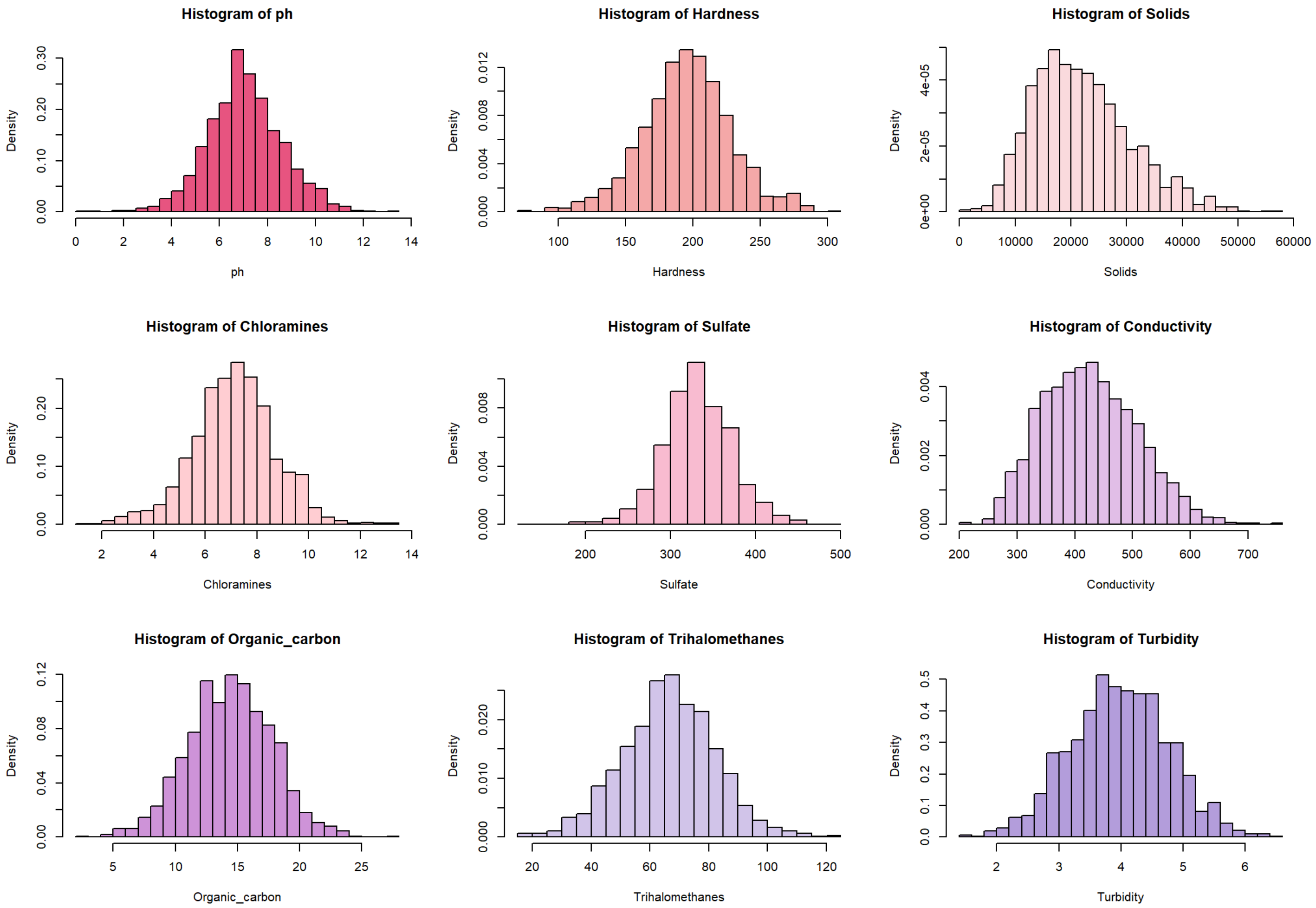
1. Exploratory Data Analysis (EDA)

```
##                     mean      sd    min      max
## ph                  7.07    1.56   0.23    13.35
## Hardness          195.81   32.74  73.49   306.63
## Solids          22070.01 8720.07 320.94 56488.67
## Chloramines         7.13    1.60   1.39    13.13
## Sulfate           332.10   40.85 129.00   481.03
## Conductivity      426.43   81.76 201.62   753.34
## Organic_carbon     14.41    3.37   2.20    27.01
## Trihalomethanes    66.48   15.82  15.68   124.00
## Turbidity           3.98    0.79   1.45     6.49
```

Our first step is to analyze the distribution of the potability variable through a bar plot



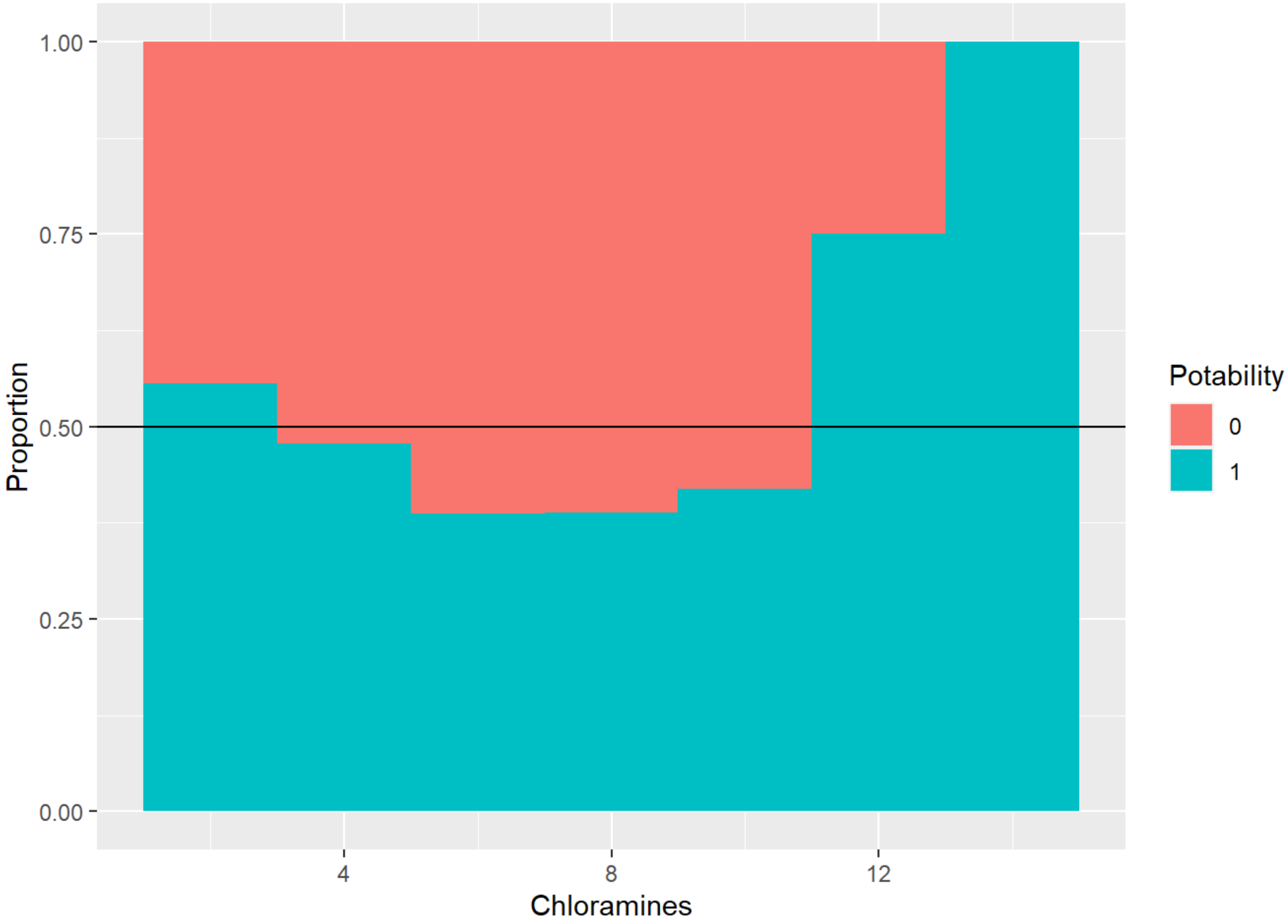
0
200
400
600
800
1000
Not Potable
Potable



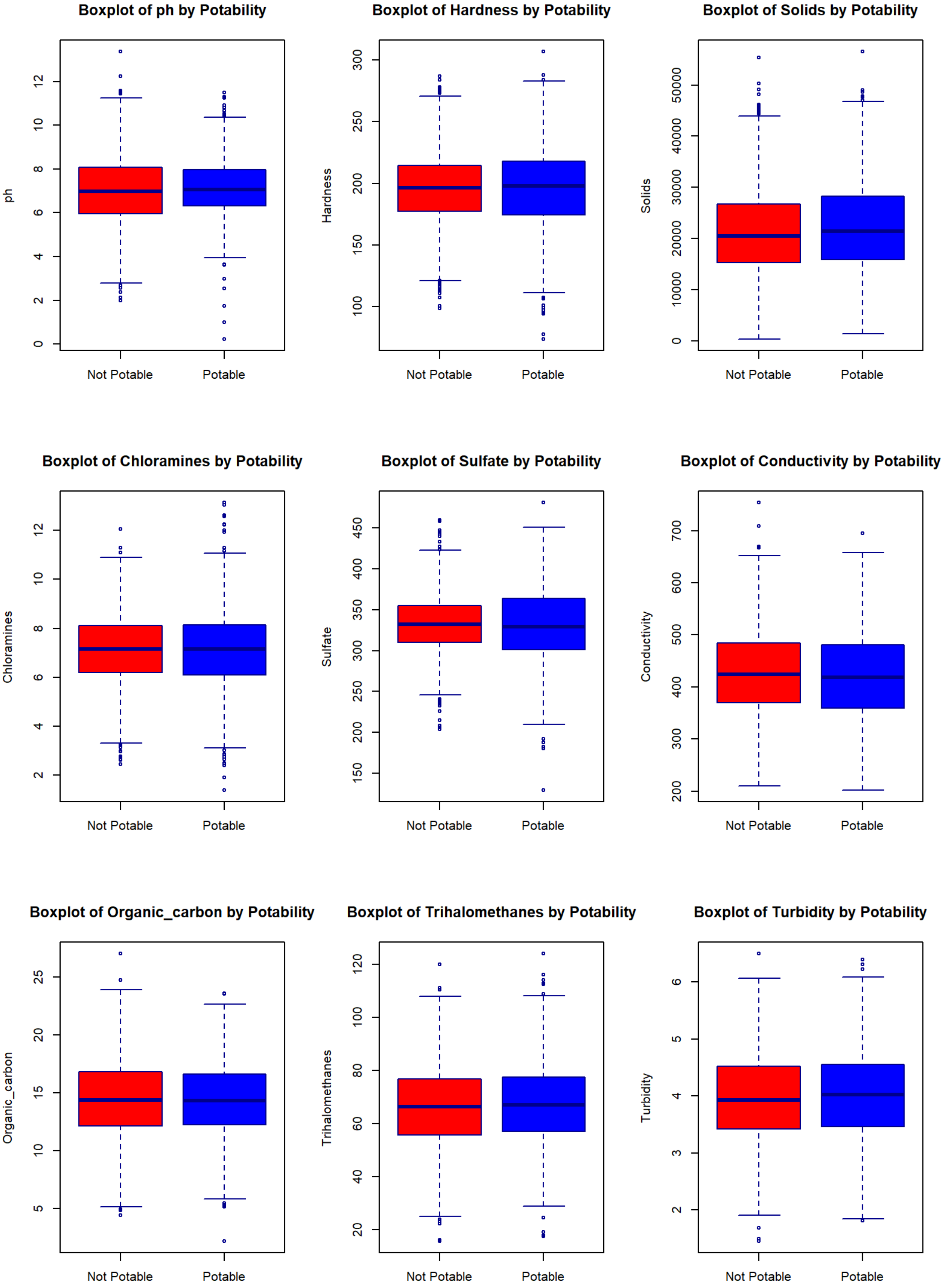
Histogram of ph
Density
0.00 0.10 0.20 0.30
0 2 4 6 8 10 12 14
ph
Histogram of Hardness
Density
0.000 0.004 0.008 0.012
100 150 200 250 300
Hardness
Histogram of Solids
Density
0e+00 2e-05 4e-05
0 10000 20000 30000 40000 50000 60000
Solids
Histogram of Chloramines
Density
0.00 0.10 0.20
2 4 6 8 10 12 14
Chloramines
Histogram of Sulfate
Density
0.000 0.004 0.008
200 300 400 500
Sulfate
Histogram of Conductivity
Density
0.000 0.002 0.004
200 300 400 500 600 700
Conductivity
Histogram of Organic_carbon
Density
0.00 0.04 0.08 0.12
5 10 15 20 25
Organic_carbon
Histogram of Trihalomethanes
Density
0.000 0.010 0.020
20 40 60 80 100 120
Trihalomethanes
Histogram of Turbidity
Density
0.0 0.1 0.2 0.3 0.4 0.5
2 3 4 5 6
Turbidity

These are the histograms related to all the variables in our dataset. We also tried to compute the histogram for each variable based on their potability status, but we could not derive any meaningful information from it.

Looking at our dataset, chloramine emerged as a potentially significant variable, particularly for values exceeding 10, where we observed high levels of potability. This suggested a possible correlation worth further investigation.



However, upon conducting additional tests and analyzing the histograms, we found that the observed differences were primarily due to discrepancies in the data distribution. Consequently, no conclusive evidence supported a strong or meaningful relationship. The results remained statistically insignificant, indicating that chloramine alone may not be a determining factor in water potability.



Boxplot of ph by Potability

Boxplot of Hardness by Potability

Boxplot of Solids by Potability

Boxplot of Chloramines by Potability

Boxplot of Sulfate by Potability

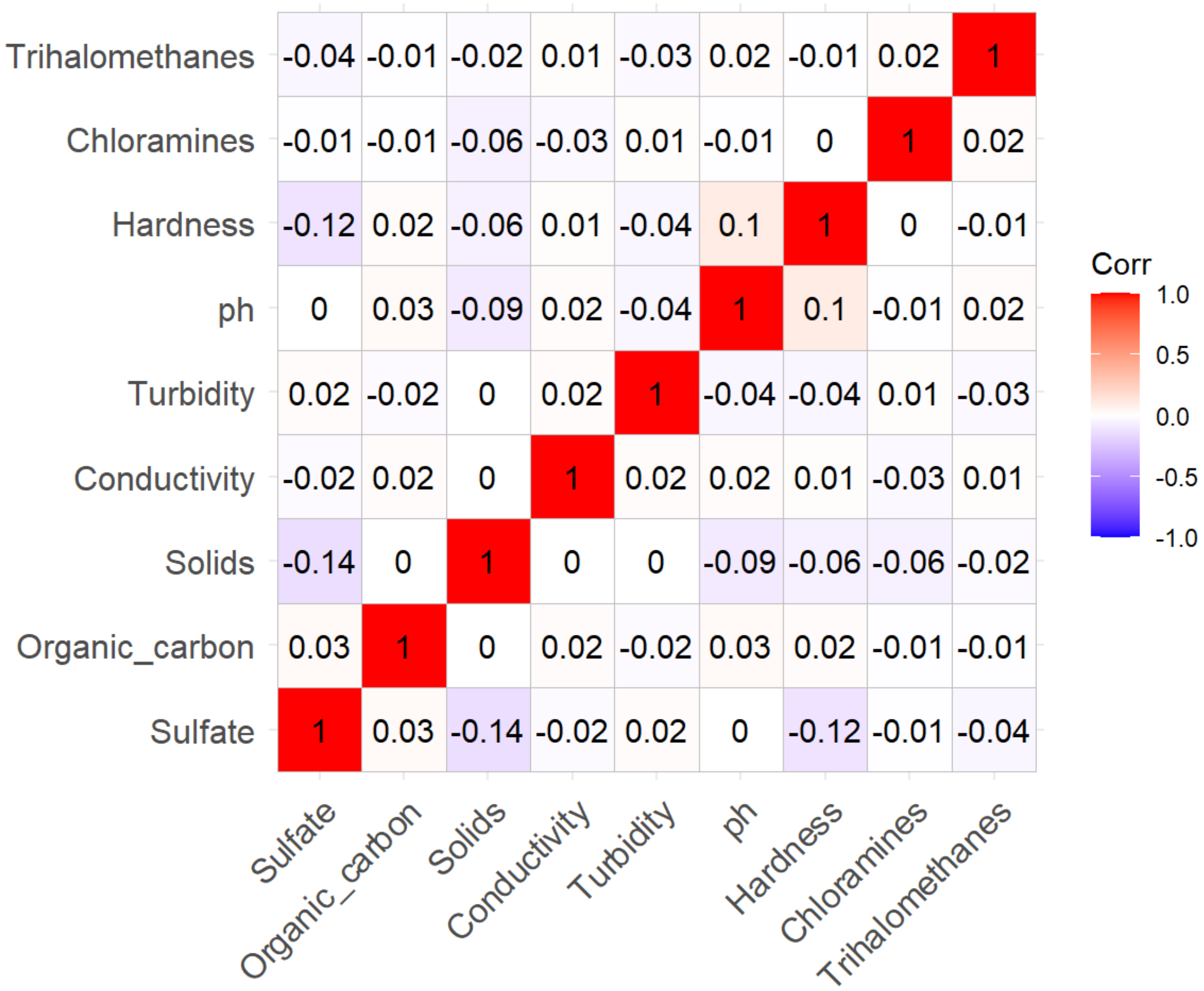
Boxplot of Conductivity by Potability

Boxplot of Organic_carbon by Potability

Boxplot of Trihalomethanes by Potability

Boxplot of Turbidity by Potability

As the last step of our EDA, we compute the correlation matrix, which, like the previous techniques, doesn't show significant results



Conclusion

As we can see from the scatterplot, the boxplot, and the correlation matrix, there is no clear linear correlation between the variables, nor is there a distinct separation between the two groups defined by the response variable (potable and non-potable water). Additionally, we observed that the dataset is artificially generated based on a certain distribution. When comparing some variable values with real-world data, inconsistencies become apparent—for instance, there are unusually low pH values which, in reality, would indicate non-potable water, yet in our dataset, these samples are labeled as potable. This suggests that the data may not accurately reflect real-world conditions.

2. Fitting the models

```
##
##   0   1
## 143  99
```

First of all, we splitted the dataset in two parts, in order to generate a validation set, then we tried to find the best model for the training data,

2.1 Logistic regression

```
## 
## Call:
## glm(formula = Potability ~ ., family = binomial, data = train_set)
## 
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)  
## (Intercept)     -1.237e+00  9.098e-01  -1.360   0.1739  
## ph               4.805e-02  3.561e-02   1.350   0.1771  
## Hardness        -8.096e-04  1.724e-03  -0.470   0.6387  
## Solids           9.262e-06  6.453e-06   1.435   0.1512  
## Chloramines      5.733e-03  3.491e-02   0.164   0.8696  
## Sulfate         -7.484e-04  1.394e-03  -0.537   0.5915  
## Conductivity    -5.314e-04  6.841e-04  -0.777   0.4373  
## Organic_carbon  -4.164e-03  1.664e-02  -0.250   0.8024  
## Trihalomethanes  3.654e-03  3.487e-03   1.048   0.2947  
## Turbidity        1.769e-01  7.018e-02   2.520   0.0117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1840.8  on 1365  degrees of freedom
## Residual deviance: 1828.9  on 1356  degrees of freedom
## AIC: 1848.9
## 
## Number of Fisher Scoring iterations: 4
```

The logistic regression model shows that Turbidity is the only statistically significant variable, with a p-value of 0.0117, suggesting it positively influences water potability, while the other variables are not significant. The model improved slightly over the null model, as indicated by the reduction in residual deviance (1828.9 vs. 1840.8), but the overall fit remains modest. To simplify the model, we also tested a version with only Turbidity as a predictor, but the performance did not improve significantly. Finally, adjusting the classification threshold did not lead to a better predictions, further supporting the limited predictive power of the model.

```
##
## Call:
## glm(formula = Potability ~ Turbidity, family = binomial, data = train_set)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.06942    0.28480  -3.755 0.000173 ***
## Turbidity    0.16847    0.06987   2.411 0.015904 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1840.8  on 1365  degrees of freedom
## Residual deviance: 1834.9  on 1364  degrees of freedom
## AIC: 1838.9
##
## Number of Fisher Scoring iterations: 4
```

The logistic regression model with Turbidity as the sole predictor shows it is significant (p-value = 0.0159), indicating a positive effect on water potability. However, the small reduction in residual deviance (1834.9 vs. 1840.8) and the AIC of 1838.9 suggest the model offers only a modest improvement.

```
##
##       0   1
##   0 116  89
##   1  27  10
```

```
## Accuracy:  0.521
```

```
## Sensitivity:  0.101
```

```
## Specificity:  0.811
```

Moreover, we tried to test the model on the validation set, with low values for both accuracy and sensitivity, so we move on to other models, like LDA.

2.2 LDA

```
## Call:
## lda(Potability ~ ., data = train_set)
##
## Prior probabilities of groups:
##         0         1 
## 0.5980966 0.4019034 
##
## Group means:
##         ph Hardness   Solids Chloramines  Sulfate Conductivity Organic_carbon
## 0 7.018067 196.2207 21803.94    7.089638 332.6496     428.8313       14.42833
## 1 7.111263 195.4243 22467.84    7.100780 331.2171     425.5752       14.37020
##   Trihalomethanes Turbidity
## 0        66.04290  3.935159
## 1        66.93948  4.041126
##
## Coefficients of linear discriminants:
##                           LD1
## ph               0.2519190825
## Hardness        -0.0042303104
## Solids           0.0000486227
## Chloramines      0.0298346658
## Sulfate         -0.0039413488
## Conductivity    -0.0027817252
## Organic_carbon  -0.0217860010
## Trihalomethanes  0.0191449633
## Turbidity        0.9258996487
```

```
##          LD1         0         1  
## 1  0.7102542 0.5660262 0.4339738 1
## 2  0.6350648 0.5695451 0.4304549 1
## 3 -1.2632736 0.6552116 0.3447884 1
## 4 -0.2916061 0.6122359 0.3877641 1
## 5  0.2051084 0.5895226 0.4104774 1
## 6  0.6169030 0.5703940 0.4296060 1
```

```
##          Actual
## Predicted   0   1 Sum
##       0   141  92 233
##       1     2   7   9
##       Sum 143  99 242
```

```
##
##       0   1
##   0 141  92
##   1   2   7
```

```
##
## Model Performance Metrics:
```

```
## Accuracy:  0.6116 ( 61.16 % )
```

```
## Sensitivity:  0.0707
```

```
## Specificity:  0.986
```

Linear Discriminant Analysis (LDA) is a technique used to classify observations into different groups by finding a linear combination of features that best separates the groups. In this case, it helps distinguish between potable and non-potable water based on various water quality metrics. Compared to the logistic regression model, LDA reinforces Turbidity as the strongest predictor, with a coefficient of 0.93. The prior probabilities of the two classes show a slight imbalance, with 60% non-potable and 40% potable. Like before, we tried to fit our model to the validation set and after trying different models with different predictors, the one which gave us the best performance is the one with all of them. The model is performing decently in terms of accuracy (61.36%) and is better at identifying non-potable water (specificity of 60.5%). However, it struggles to identify potable water, as shown by the low True Positives (TP = 7) and the relatively low Sensitivity (77.78%). There's a significant number of False Positives (92), meaning that the model tends to incorrectly classify non-potable water as potable.

2.3 QDA

```
## Call:
## qda(Potability ~ ., data = train_set)
##
## Prior probabilities of groups:
##         0         1
## 0.5980966 0.4019034
##
## Group means:
##         ph Hardness   Solids Chloramines  Sulfate Conductivity Organic_carbon
## 0 7.018067 196.2207 21803.94    7.089638 332.6496     428.8313       14.42833
## 1 7.111263 195.4243 22467.84    7.100780 331.2171     425.5752       14.37020
##   Trihalomethanes Turbidity
## 0        66.04290  3.935159
## 1        66.93948  4.041126
## Confusion Matrix and Statistics
##
##          Actual
## Predicted   0   1
##         0 131  59
##         1  12  40
##
##                Accuracy : 0.7066
##                  95% CI : (0.6449, 0.7632)
##     No Information Rate : 0.5909
##     P-Value [Acc > NIR] : 0.0001255
##
##                   Kappa : 0.3453
##
##  Mcnemar's Test P-Value : 4.783e-08
##
##             Sensitivity : 0.9161
##             Specificity : 0.4040
##          Pos Pred Value : 0.6895
##          Neg Pred Value : 0.7692
##              Prevalence : 0.5909
##          Detection Rate : 0.5413
##    Detection Prevalence : 0.7851
##       Balanced Accuracy : 0.6601
##
##        'Positive' Class : 0
##
```

Like the LDA, we found that also in this case, the model with all the predictors is the best one

```
## Call:
## qda(Potability ~ Hardness + Chloramines + Sulfate + Trihalomethanes,
##     data = train_set)
##
## Prior probabilities of groups:
##         0         1
## 0.5980966 0.4019034
##
## Group means:
##   Hardness Chloramines  Sulfate Trihalomethanes
## 0 196.2207    7.089638 332.6496        66.04290
## 1 195.4243    7.100780 331.2171        66.93948
```

```
##
##       0   1
##   0 130  66
##   1  13  33
```

```
##
## QDA Model Performance Metrics:
```

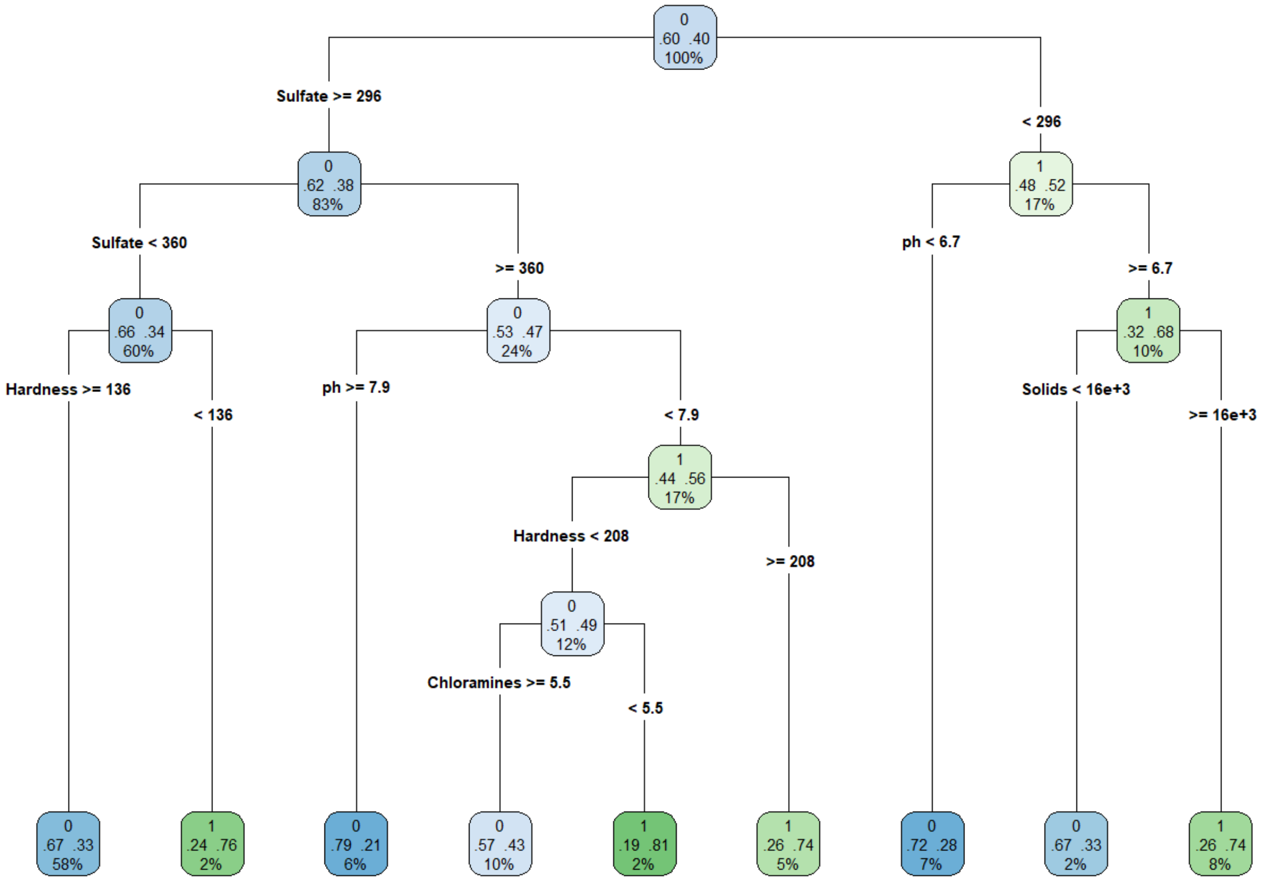
```
## Accuracy:     0.6736 ( 67.36 % )
```

```
## Sensitivity:  0.3333
```

```
## Specificity:  0.9091
```

The second best QDA (Quadratic Discriminant Analysis) model we fitted included only four predictors, sacrificing just 3% in performance for greater simplicity. However, the model still underperforms, with the following key metrics: Accuracy is 67.36%, indicating moderate overall performance; Sensitivity is 33.33%, reflecting poor detection of potable water; and Specificity is 90.91%, demonstrating strong identification of non-potable water.

2.4 Tree



```
##
## Training Set Performance:
```

```
##       Predicted
## Actual   0   1
##      0 758  59
##      1 371 178
## [1] "Accuracy: 0.6852"
```

```
## [1] "Sensitivity (Recall): 0.3242"
```

```
## [1] "Specificity: 0.9278"
```

```
##
## Validation Set Performance:
```

```
##       Predicted
## Actual   0   1
##      0 136   7
##      1  67  32
```

```
## [1] "Accuracy: 0.6942"
```

```
## [1] "Sensitivity (Recall): 0.3232"
```

```
## [1] "Specificity: 0.951"
```

Since the parametric approach (QDA) did not fit the data well, we switched to a non-parametric model—a decision tree—to see if we could achieve better results. We first grew a full tree and then pruned it to avoid overfitting while maintaining interpretability. The final model performed well in terms of specificity (95.10%), meaning it was excellent at correctly identifying non-potable water. However, it struggled with accuracy and sensitivity, indicating poor detection of potable water cases. Overall, the tree model did not generalize well to new data, which likely explains its poor performance when tested on the validation set.

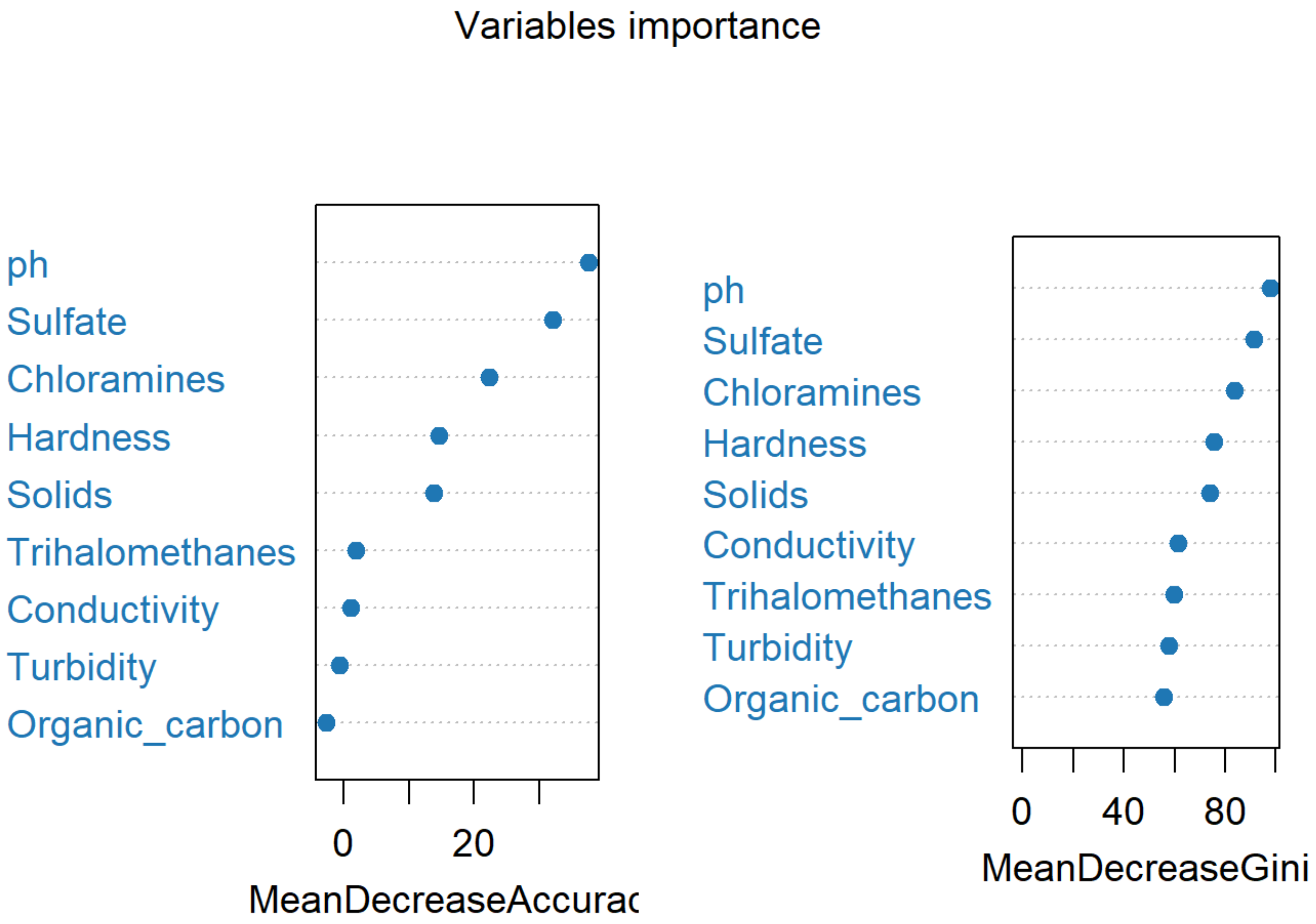
2.5 Random Forest

```
##           Reference
## Prediction   0   1
##          0 120  53
##          1  23  46
```

```
##
## Accuracy: 0.686
```

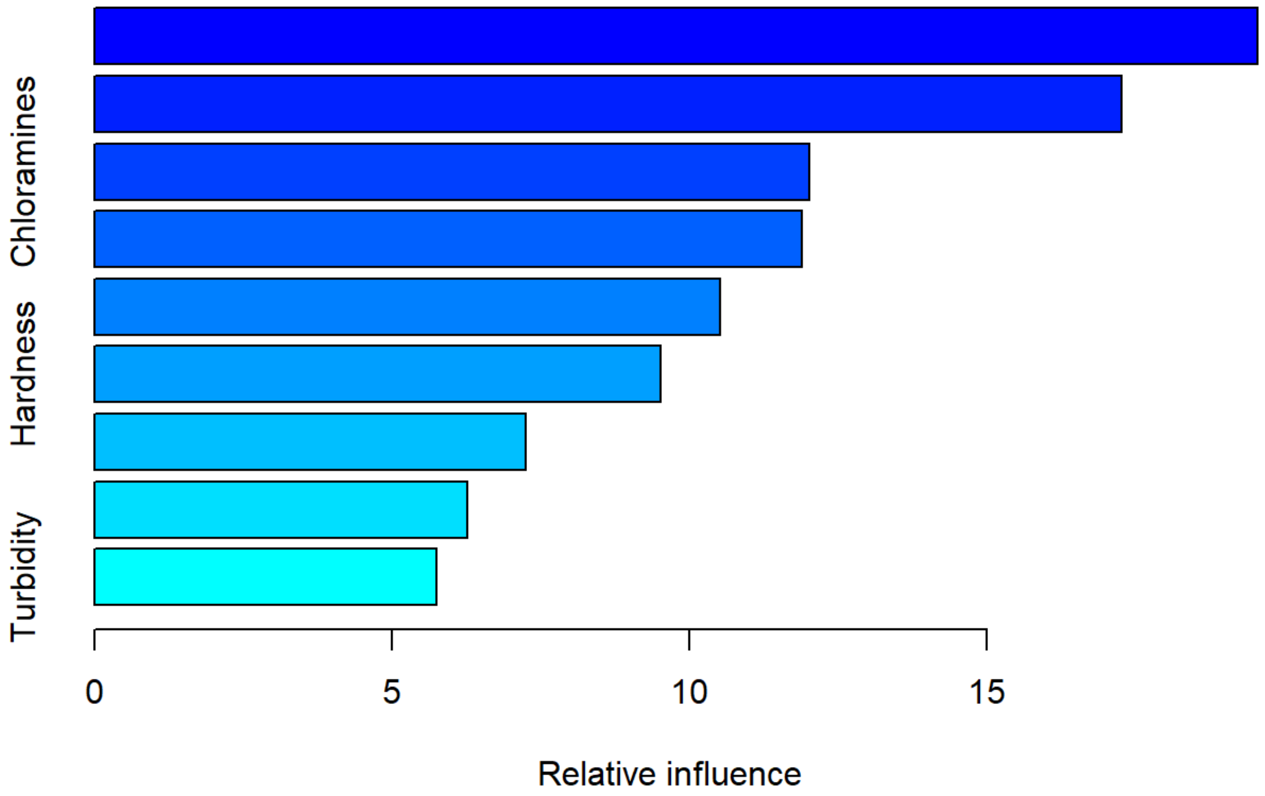
```
##
## Sensitivity (Recall): 0.839
```

```
##
## Specificity: 0.465
```



We attempted to model the data using a Random Forest classifier with the bagging technique. This approach demonstrated high sensitivity (83.9%) but low specificity (46.5%). In other words, the model is effective at identifying non-potable water (class 0) but struggles to correctly detect potable water (class 1). This suggests a bias toward classifying water as non-potable, likely to minimize false positives, but at the cost of overlooking many true potable cases. By using Random Forest with bagging (500 trees and the full feature set considered at each split), we improved the model's stability and reduced variance compared to a single decision tree. The most influential features in predicting water potability were: pH, Sulfate, Chloramines, Hardness, Solids. Despite these insights, the model achieved only moderate accuracy (~68.6%), which is similar to the performance of the simpler decision tree model. This limited improvement may be attributed to class imbalance or the inherent complexity of water quality patterns.

Random forest with boosting technique:



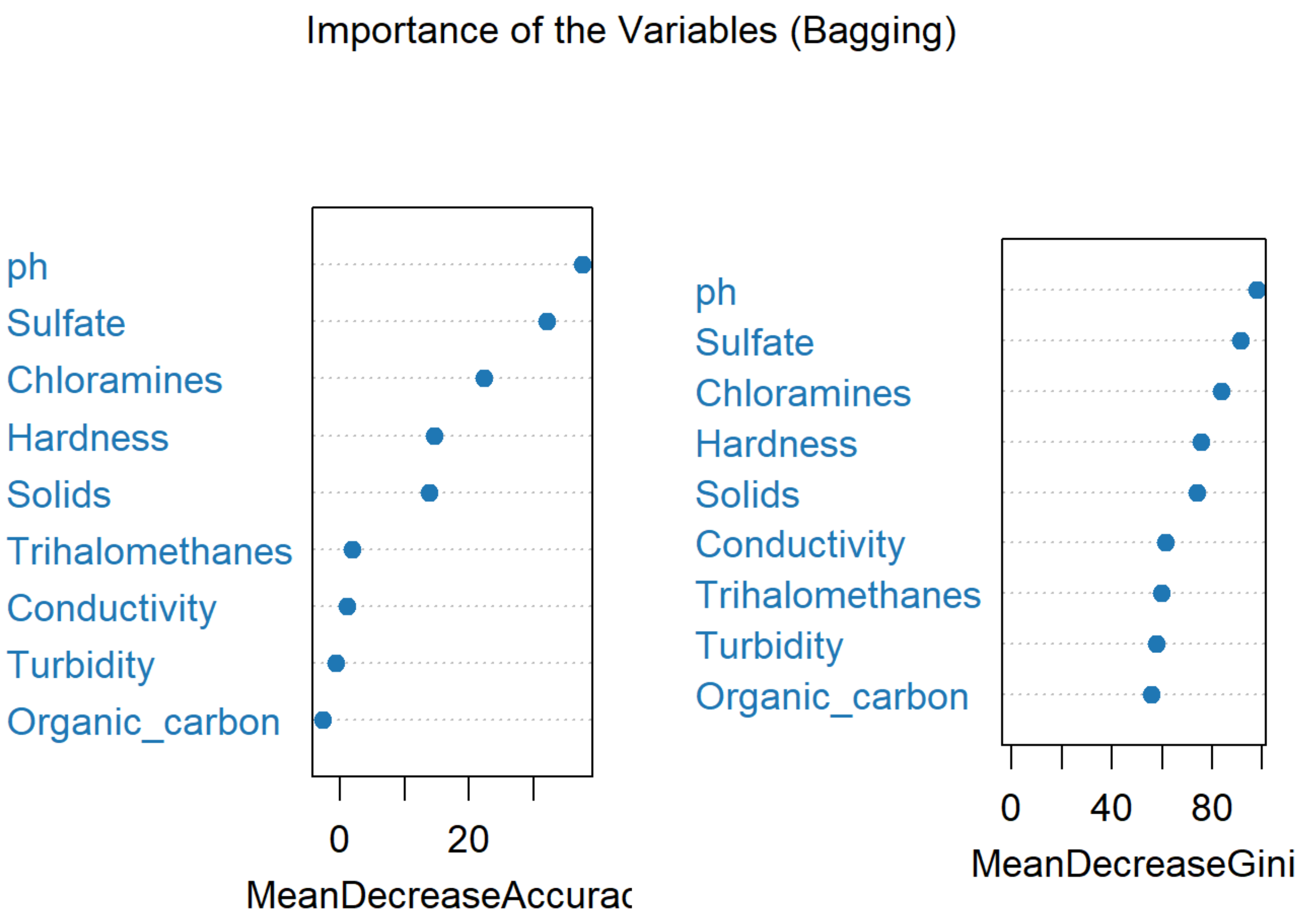
```
##                              var   rel.inf
## ph                            ph 19.540397
## Sulfate                  Sulfate 17.257317
## Chloramines          Chloramines 12.015616
## Solids                    Solids 11.891085
## Conductivity        Conductivity 10.514629
## Hardness                Hardness  9.516968
## Trihalomethanes Trihalomethanes  7.251034
## Organic_carbon    Organic_carbon  6.257113
## Turbidity              Turbidity  5.755840
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 116  49
##          1  27  50
##                                           
##                Accuracy : 0.686           
##                  95% CI : (0.6234, 0.7439)
##     No Information Rate : 0.5909          
##     P-Value [Acc > NIR] : 0.001434        
##                                           
##                   Kappa : 0.3274          
##                                           
##  Mcnemar's Test P-Value : 0.016002        
##                                           
##             Sensitivity : 0.8112          
##             Specificity : 0.5051          
##          Pos Pred Value : 0.7030          
##          Neg Pred Value : 0.6494          
##              Prevalence : 0.5909          
##          Detection Rate : 0.4793          
##    Detection Prevalence : 0.6818          
##       Balanced Accuracy : 0.6581          
##                                           
##        'Positive' Class : 0               
## 
```

Because we got bad results, we also tried the random forest with the boosting technique, but we still got poor results, so we changed the approach.

Balancing the data set with library rose:

```
## 
##   0   1 
## 841 767
```

We used the library Random Over-Sampling Examples (ROSE) generates new synthetic data for the minority class (oversampling) and/or reduces the majority class (undersampling) to mitigate the problem of class imbalance and improve the performane of the model


Importance of the Variables (Bagging)
ph
Sulfate
Chloramines
Hardness
Solids
Trihalomethanes
Conductivity
Turbidity
Organic_carbon
0
20
MeanDecreaseAccurac
ph
Sulfate
Chloramines
Hardness
Solids
Conductivity
Trihalomethanes
Turbidity
Organic_carbon
0
40
80
MeanDecreaseGini

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 122  38
##          1  21  61
##                                           
##                Accuracy : 0.7562          
##                  95% CI : (0.6971, 0.8089)
##     No Information Rate : 0.5909          
##     P-Value [Acc > NIR] : 5.002e-08       
##                                           
##                   Kappa : 0.482           
##                                           
##  Mcnemar's Test P-Value : 0.03725         
##                                           
##             Sensitivity : 0.8531          
##             Specificity : 0.6162          
##          Pos Pred Value : 0.7625          
##          Neg Pred Value : 0.7439          
##              Prevalence : 0.5909          
##          Detection Rate : 0.5041          
##    Detection Prevalence : 0.6612          
##       Balanced Accuracy : 0.7347          
##                                           
##        'Positive' Class : 0               
## 
```

Using ROSE and applying the bagging technique to our dataset, we achieved the best accuracy of 75.6% among all models tested, with high sensitivity (85.3%) and moderate specificity (61.6%).