ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA KỸ THUẬT HÓA HỌC

BÁO CÁO BÀI TẬP LỚN

MÔN

XÁC XUẤT VÀ THỐNG KÊ

MT2013

Xác định ảnh hưởng của một số thành phần trong nước đến khả năng uống được trên phạm vi toàn thế giới thông qua nghiên cứu thống kê mẫu

L13_NHÓM 14

Giảng viên hướng dẫn: TS. Phan Thị Hường

Thành phố Hồ Chí Minh, tháng 11, năm 2023

DANH SÁCH THÀNH VIÊN NHÓM

STT	Mã số SV	Họ và tên	Mô tả đóng góp	Điểm chia
1	2212417	Bùi Thị Tuyết Nhi	Phần I, III	-0,5
2	2212625	Nguyễn Hoàng Phúc	Phần I, VII	-0,5
3	2212702	Tạ Thị Trúc Phương	Phần IV	-0,5
4	2212772	Cao Anh Quân	Phần V	+1,5

MỤC LỤC

I. TỔNG QUAN DỮ LIỆU

1.1. Giới thiệu chung

Tiếp cận nguồn nước uống an toàn là điều thiết yếu cho sức khỏe cũng như là quyền cơ bản của con người và cũng góp một phần vào chính sách bảo vệ sức khỏe hiệu quả. Đây là vấn đề quan trọng về sức khỏe và phát triển ở cấp quốc gia, khu vực và địa phương. Ở một số vùng, người ta đã chứng minh rằng đầu tư vào nguồn cung cấp và vệ sinh nước có thể mang lại lợi ích kinh tế ròng vì việc giảm các ảnh hưởng tiêu cực đến sức khỏe và chi phí chăm sóc sức khỏe sẽ có ảnh hưởng lớn hơn chi phí thực hiện các biện pháp can thiệp.

Trong đề tài này, ta sẽ phân tích tệp dữ liệu chứa các chỉ số liên quan đến khả năng tiêu thụ nước của 3276 mẫu nước khác nhau được thu thập tại nhiều vùng khắc nhau trên thế giới.

Dữ liệu đầu vào có chứa tệp lưu trữ dưới dạng CSV: water_potability.csv cho các mẫu nước đã được thu thập. Bảng dữ liệu có danh sách các cột biểu thị cho thông tin cần thiết của dữ liệu và một số đặc đặc trưng sẽ bao gồm: giá trị pH, độ cứng của nước, nồng độ chloramines, sulfate, cacbon hữu cơ và trihalomethanes, độ dẫn điện, tổng chất rắn hòa tan và độ đục của nước.

Dữ liệu được trích từ:

https://www.kaggle.com/datasets/adityakadiwal/water-potability/

1.2. Các biến trong bảng dữ liệu

STT	Tên biến	Loại biến	Tập giá trị	Mô tả
1	pH	Liên tục	$\{x \in R \mid 0 \leq x \leq 14\}$	Thông số quan trọng trong việc đánh giá cân bằng axit-bazo trong nước.
2	Hardness (mg/L)	Liên tục	$\{x \in R \mid 47,4 \leq x \leq 323\}$	Khả năng của nước làm kết tủa xà phòng do Canxi và Magie gây ra
3	Solids (ppm)	Liên tục	$\{x \in R \mid 321 \leq x \leq 61,2 \times 10^3\}$	Chỉ số thể hiện tổng chất rắn hòa tan tồn tại trong một thể tích nước nhất định. (canxi, magiê, natri, kali và các anion cacbonat, bicarbonate, …)
4	Chloramines (ppm)	Liên tục	$\{x \in R \mid 0,35 \leq x \leq 13,1\}$	Chất khử trùng chính được sử dụng trong hệ thống nước công cộng
5	Sulfate (mg/L)	Liên tục	$\{x \in R \mid 129 \leq x \leq 481\}$	Sulfates là những chất xuất hiện tự nhiên được tìm thấy trong khoáng chất, đất và đá. Chúng có mặt trong không khí xung quanh, nước ngầm, thực vật và thực phẩm.
6	Conductivity (μS/cm)	Liên tục	$\{x \in R \mid 181 \leq x \leq 753\}$	Độ dẫn điện (EC) thực chất đo quá trình ion của dung dịch cho phép nó truyền dòng điện

7	Organic_carbon (ppm)	Liên tục	$\{x \in R \mid 2,2 \leq x \leq 28,3\}$	Là thước đo tổng lượng cacbon có trong các hợp chất hữu cơ trong nước tinh khiết.
8	Trihalomethanes (µg/L)	Liên tục	$\{x \in R \mid 0,74 \leq x \leq 124\}$	Là những hóa chất có thể tìm thấy trong nước được xử lý bằng clo.
9	Turbidity (NTU)	Liên tục	$\{x \in R \mid 1,45 \leq x \leq 6,74\}$	Độ đục của nước phụ thuộc vào lượng chất rắn có ở trạng thái lơ lửng. Nó là thước đo đặc tính phát sáng của nước và thử nghiệm được sử dụng để chỉ ra chất lượng xả thải đối với chất keo
10	Potability	Rời rạc	$x = 0$ hoặc $x = 1$	Cho biết nước có an toàn cho con người hay không (1=an toàn; 0= không an toàn

II. KIẾN THỨC NỀN

2.1. Hồi quy Logistic

2.1.1. Định nghĩa

Hồi quy logistic (hoặc hồi quy logit) là một quá trình ước tính xác suất của một kết quả riêng biệt, dựa trên một tập dữ liệu nhất định gồm các biến độc lập. Nó là một phương pháp phân tích hồi quy để tiến hành khi biến phụ thuộc là nhị phân (có giá trị 0 hoặc 1).

Giống như tất cả các phân tích hồi quy, hồi quy logistic là một phương pháp phân tích mang tính dự đoán. Hồi quy logistic được sử dụng để mô tả dữ liệu và giải thích mối quan hệ giữa một dữ liệu phụ thuộc có biến nhị phân vào một hoặc nhiều biến độc lập. Kỹ thuật hồi quy này gần tương tự như hồi quy tuyến tính và có thể được sử dụng để dự đoán các vấn đề xác suất phân loại

Trong hồi quy logistic, phép biến đổi logit được áp dụng theo tỷ lệ cược—nghĩa là xác suất thành công chia cho xác suất thất bại. Điều này cũng thường được gọi là tỷ lệ cược log hoặc logarit tự nhiên của tỷ lệ cược và hàm logistic này được biểu diễn bằng công thức sau:

$$P(X) = \frac{1}{1 + e^{-\beta x}}$$

Trong đó:

- P: "Xác suất thành công" - xác suất của biến phụ thuộc bằng thành công/ có xảy ra chứ không phải là thất bại/không xảy ra (xác suất là 1)

- $X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$: là biến phụ thuộc
- k: số lượng tham số
- X_i: Biến độc lập
- β_0: Hệ số chặn
- β_i: Hệ số của X_i

Với một biến X, mô hình lý thuyết cho P có dạng tín hiệu kéo dài với các tiệm cận tại 0 và 1, mặc dù trong ước tính mẫu chúng ta có thể không thấy được hình dạng được đề cập nếu phạm vi của biến bị giới hạn.

Sau khi mô hình đã được tính toán, cách tốt nhất là đánh giá mức độ hiệu quả của mô hình dự đoán biến phụ thuộc, được gọi là mức độ phù hợp. Kiểm định Hosmer–Lemeshow là một phương pháp phổ biến để đánh giá mức độ phù hợp của mô hình, sẽ được thảo luận trong phần tiếp theo.

2.2.2. Hồi quy logistic vs Hồi quy tuyến tính

Mô hình hồi quy tuyến tính được sử dụng để xác định mối quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Khi chỉ có một biến độc lập và một biến phụ thuộc, nó được gọi là hồi quy tuyến tính đơn giản, nhưng khi số lượng biến độc lập tăng lên, nó được gọi là hồi quy tuyến tính bội. Đối với mỗi loại hồi quy tuyến tính, nó tìm cách vẽ một đường cong phù hợp nhất thông qua một tập hợp các điểm dữ liệu, thường được tính bằng phương pháp bình phương cực tiểu.

Tương tự như hồi quy tuyến tính, mô hình hồi quy logistic cũng được sử dụng để ước lượng mối quan hệ giữa một biến phụ thuộc vào một hoặc nhiều biến độc lập, nhưng nó được sử dụng để đưa ra dự đoán về một biến phân loại so với một biến liên tục. Mô hình sẽ mang lại kết quả nhị phân hoặc chia đôi giới hạn ở hai kết quả có thể xảy ra: có/không, 0/1, hoặc đúng/sai. Đơn vị đo cũng khác với hồi quy tuyến tính vì nó tạo ra kết quả là một xác suất, nhưng hàm logit biến đường cong S thành một đường thẳng.

Với công dụng của nó để giải quyết các bài toán phân loại, hồi quy logistic sẽ được sử dụng trong đề tài này nhằm xác định xác suất uống được của các mẫu nước bằng cách xác định mối quan hệ giữa các biến số như độ pH, độ dẫn điện, độ cứng v.v. của một mẫu nước và sử dụng nó để dự đoán liệu mẫu nước đó có đủ an toàn cho việc sử dụng trong việc uống hay không.

2.2. Kiểm định Hosmer-Lemeshow

Hiệu suất tổng thể về mức độ phù hợp của mô hình có thể được đo bằng một số thử nghiệm mức độ phù hợp khác nhau, một trong số đó là thử nghiệm Hosmer-Lemeshow. Đây là bài kiểm tra mức độ phù hợp của hồi quy logistic, đặc biệt đối với các mô hình dự đoán rủi ro, cho biết dữ liệu tốt đến mức nào và có phù hợp với mô hình hay không. Về cơ bản, đây là phép thử mức độ phù hợp của chi square cho dữ liệu được nhóm và được tiến hành bằng cách sắp xếp n bản ghi trong tập dữ liệu theo ước tính xác suất thành công, chia tập hợp đã sắp xếp thành g nhóm có kích thước bằng nhau và đánh giá thống kê Hosmor-Lemeshow:

$$\widehat{C_g} = \sum_{i=1}^{g} \left[\frac{\left(O_{s,i} - E_{s,i}\right)^2}{E_{s,i}} + \frac{\left(O_{f,i} - E_{f,i}\right)^2}{E_{f,i}} \right]$$

- $O_{s,i}$: số lần diễn ra thành công quan sát được

- $O_{f,i}$: số lần diễn ra thất bại quan sát được

- $E_{s,i}$: số lần được đoán sẽ xảy ra thành công ở nhóm thứ i

- $E_{f,i}$: số lần được dự đoán sẽ xảy ra thất bại ở nhóm thứ i

Giả thuyết không và đối thuyết của thử nghiệm là:

$$\begin{cases} H_0 : \ O_{s,i} = E_{s,i} \\ H_1 : O_{s,i} \neq E_{s,i} \end{cases}$$

Giả thuyết còn có thể được hiểu là:

$$\begin{cases} H_0 : Mô\ hình\ phù\ hợp\ với\ dữ\ liệu \\ H_1 : Mô\ hình\ không\ phù\ hợp\ với\ dữ\ liệu \end{cases}$$

Theo giả thuyết không rằng mô hình phù hợp với dữ liệu, chúng tôi chứng minh rằng $\widehat{C_g}$ có phân phối χ^2 với g−2 bậc tự do. Do đó, p-value của kiểm định Hosmer-Lemeshow là:

$$\int_{\widehat{C_g}}^{\infty} \chi^2_{g-2}(x)dx$$

Với χ^2_{g-2}(x) là hàm mật độ xác suất của phân phối χ^2 với g-2 bậc tự do tại x. Giá trị g được người sử dụng quyết định, nhưng giá trị thường được sử dụng là g = 10, đây là giá trị mặc định được sử dụng bởi hầu hết các mô hình thống kê.

Kiểm định về mức độ phù hợp Hosmer-Lemeshow rất hữu ích cho các tập dữ liệu không lặp lại hoặc cho các tập dữ liệu chỉ chứa một vài quan trắc được lặp lại, trong khi các thử nghiệm khác như kiểm định Pearson về mức độ phù hợp chi bình phương và bài kiểm tra mức độ phù hợp về độ lệch chuẩn cần có dữ liệu được lặp lại.

Bài kiểm định này thường được thực hiện trên máy tính nên phù hợp với đề tài này. Các kết quả được đưa ra là giá trị chi bình phương và p-value. Giá trị p nhỏ có nghĩa là mô hình phù hợp kém

2.3. Hệ số tương quan Pearson

Còn được biết với tên gọi là r của Pearson, hoặc được biết rộng rãi hơn là hệ số tương quan, hệ số tương quan Pearson là thước đo mối quan hệ tuyến tính giữa hai bô dữ liệu. Nó là tỉ số giữa hiệp phương sai của hai biến và tích của độ lệch chuẩn của chúng.

Khi áp dụng hệ số tương quan Pearson vào một mẫu thử, nó thường được biểu diễn bởi r_{xy}. Khi có các cặp dữ liệu (x_1, y_1), …, (x_n,y_n) gồm n cặp, r_{xy} có thể được xác định bởi công thức sau:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

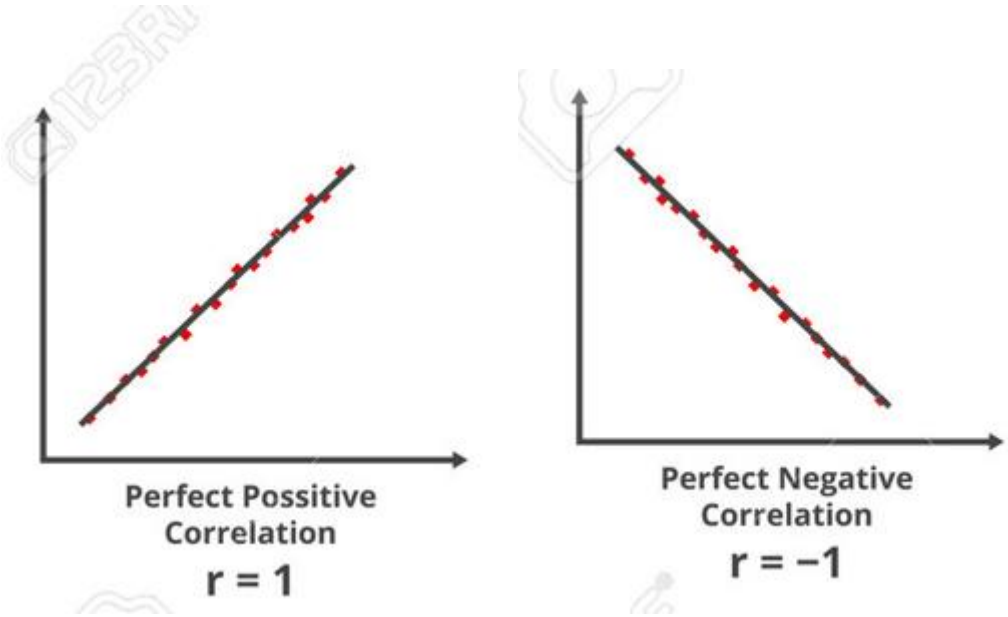
Trong đó:

- n: kích thước mẫu

-x_i, y_i: giá trị các điểm ở mẫu thứ i

- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: trung bình mẫu

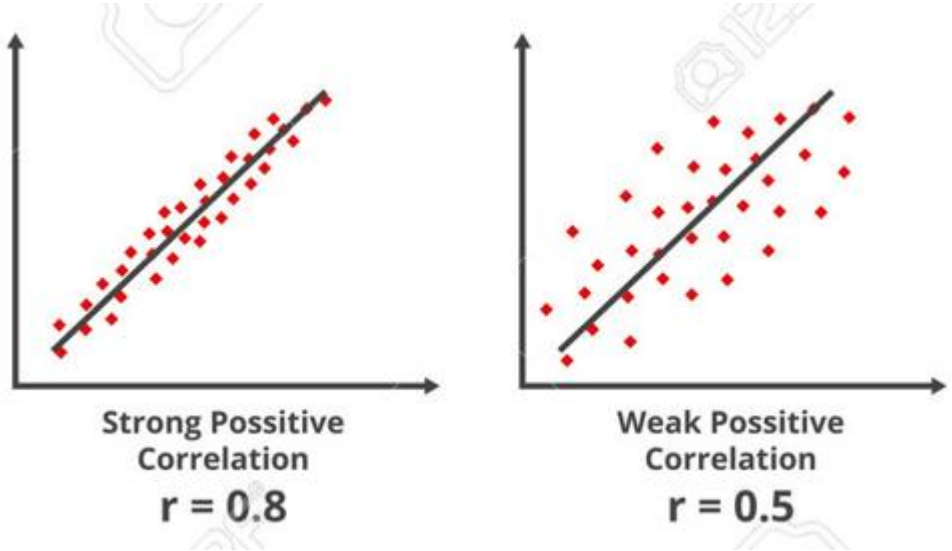
Hệ số tương quan Pearson có tính đối xứng: Cor(X,Y) = Cor(Y,X). Các giá trị của hệ số tương quan Pearson nằm trong khoảng −1 và 1, |r| = 1 có nghĩa là có một phương trình tuyến tính mô tả mối quan hệ giữa X và Y một cách hoàn hảo, với tất cả dữ liệu các điểm nằm chính xác trên một đường thẳng, trong khi khi r gần bằng 0 thì chỉ ra rằng các điểm đó nằm cách xa đường thẳng tuyến tính biểu diễn mối quan hệ giữa 2 biến X và Y:



Hình 1: Tương quan dương hoàn hảo (r = 1) và tương quan âm hoàn hảo (r = −1)

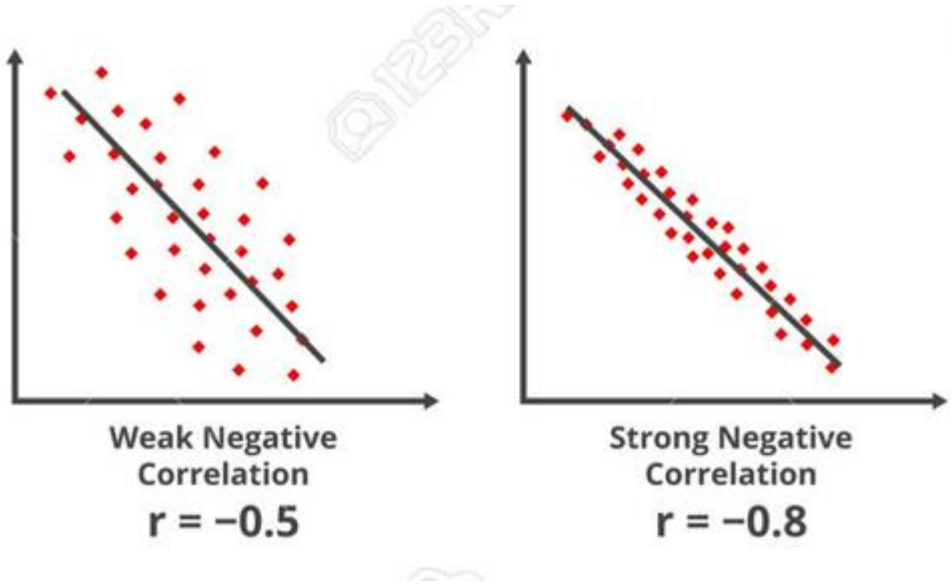
Dấu của hệ số tương quan được xác định bởi độ dốc của mô hình hồi quy:

- Hệ số tương quan có giá trị giữa 0 và 1 được gọi là tương quan dương và nó được hiểu là khi một biến thay đổi, biến kia thay đổi theo cùng hướng với biến thay đổi.



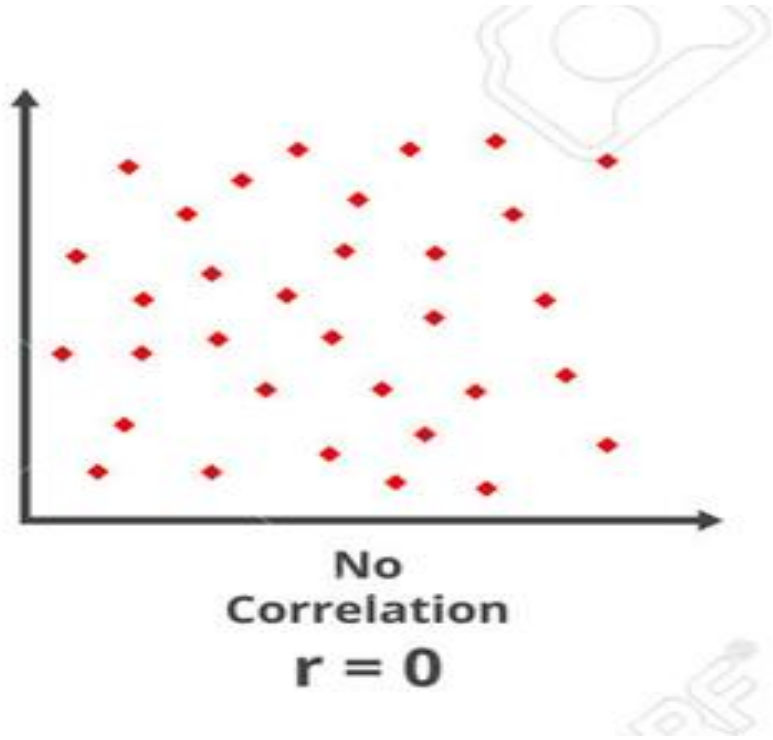
Hình 2: Hệ số tương quan dương mạnh và tương quan dương yếu

- Hệ số tương quan có giá trị từ -1 đến 0 được gọi là tương quan âm và nó được hiểu là khi một biến thay đổi thì biến kia thay đổi theo hướng ngược lại.



Hình 3: Hệ số tương quan âm mạnh va tương quan âm yếu

- Khi giá trị hệ số tương quan bằng 0, ta có thể chỉ ra rằng không có sự phụ thuộc tuyến tính giữa 2 biến.



Hình 4: Hệ số không có sự tương quan

III. TIỀN XỬ LÝ DỮ LIỆU

3.1. Đọc dữ liệu

Đầu tiên, chúng ta sẽ khai báo các thư viện cần thiết để sử dụng sau này: ggplot2, dplyr, plotly, cowplot, caret, vcd, ResourceSelection, pROC, corrplot.

Ta dùng lệnh read.csv để đọc dữ liệu vào và dùng lệnh head để xem dữ liệu đã có được nhập vào hoàn toàn hay chưa

```
> water_potability <- read.csv("D:/AQ/Probability and Statistics/water_potability.csv")
> head(water_potability)
        pH Hardness   Solids Chloramines  Sulfate Conductivity Organic_carbon Trihalomethanes
1       NA 204.8905 20791.32    7.300212 368.5164     564.3087      10.379783        86.99097
2 3.716080 129.4229 18630.06    6.635246       NA     592.8854      15.180013        56.32908
3 8.099124 224.2363 19909.54    9.275884       NA     418.6062      16.868637        66.42009
4 8.316766 214.3734 22018.42    8.059332 356.8861     363.2665      18.436525       100.34167
5 9.092223 181.1015 17978.99    6.546600 310.1357     398.4108      11.558279        31.99799
6 5.584087 188.3133 28748.69    7.544869 326.6784     280.4679       8.399735        54.91786
  Turbidity Potability
1  2.963135          0
2  4.500656          0
3  3.055934          0
4  4.628771          0
5  4.075075          0
6  2.559708          0
```

Hình 5: 6 dòng đầu của dữ liệu

3.2. Làm sạch dữ liệu

Ta dung lệnh sum kết hợp cùng với is.na(data) để kiểm tra số dữ liệu bị khuyết.

```
> sum(is.na(water_potability))
[1] 1434
```

Hình 6: Kết quả kiểm tra dữ liệu bị khuyết

Vì dữ liệu của chúng ta có nhiều giá trị bị khuyết, chúng ta sẽ tiếp tục kiểm tra xem số lượng cũng như tỉ lệ bị khuyết ở các biến.

```
> summary(is.na(water_potability))
     pH            Hardness        Solids         Chloramines      Sulfate
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:2785      FALSE:3276      FALSE:3276      FALSE:3276      FALSE:2495
 TRUE :491                                                       TRUE :781
 Conductivity    Organic_carbon  Trihalomethanes Turbidity       Potability
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:3276      FALSE:3276      FALSE:3114      FALSE:3276      FALSE:3276
                                 TRUE :162
> r_na <- colSums(is.na(water_potability))/3276
> print(r_na)
             pH        Hardness          Solids     Chloramines         Sulfate
     0.14987790      0.00000000      0.00000000      0.00000000      0.23840049
   Conductivity  Organic_carbon Trihalomethanes       Turbidity      Potability
     0.00000000      0.00000000      0.04945055      0.00000000      0.00000000
```

Hình 6: Kiểm tra các biến bị khuyết

Quan sát kết quả, ta có nhận xét sau:

- 491 dữ liệu thuộc biến pH bị khuyết, chiếm gần 15% tổng số dữ liệu.

- 781 dữ liệu thuộc biến Sulfate bị khuyết, chiếm gần 24% tổng số dữ liệu.

- 162 dữ liệu thuộc biến Trihalomethanes bị khuyết, chiếm chỉ khoảng 5% tổng số dữ liệu.

Phương pháp xử lý dữ liệu khuyết:

- Ta nhận thấy rằng biến Sulfate và pH bị khuyết dữ liệu tương đối, nhưng vì khả năng sử dụng nước cho việc uống nước là một dữ liệu khá nhạy cảm và phụ thuộc rất nhiều vào các chỉ số đo được với mỗi biến nên nhóm quyết định sẽ xóa bỏ các dòng có biến dữ liệu bị khuyết.

Sau khi dùng lệnh na.omit, ta lọc 1265 dòng và điền bô dữ liệu mới đã qua lọc vào thư mục có tên mới.

```
> water_potability_no_na<- na.omit(water_potability)
> head(water_potability_no_na)
          pH Hardness   Solids Chloramines  Sulfate Conductivity Organic_carbon
4   8.316766 214.3734 22018.42    8.059332 356.8861     363.2665      18.436525
5   9.092223 181.1015 17978.99    6.546600 310.1357     398.4108      11.558279
6   5.584087 188.3133 28748.69    7.544869 326.6784     280.4679       8.399735
7  10.223862 248.0717 28749.72    7.513408 393.6634     283.6516      13.789695
8   8.635849 203.3615 13672.09    4.563009 303.3098     474.6076      12.363817
10 11.180284 227.2315 25484.51    9.077200 404.0416     563.8855      17.927806
   Trihalomethanes Turbidity Potability
4        100.34167  4.628771          0
5         31.99799  4.075075          0
6         54.91786  2.559708          0
7         84.60356  2.672989          0
8         62.79831  4.401425          0
10        71.97660  4.370562          0
```

Hình 7: Kiểm tra dữ liệu sau khi lọc dữ liệu khuyết

3.3. Tóm tắt dữ liệu

Đầu tiên ta tóm tắt lại những dữ liệu thống kê của bộ dữ liệu mới sau khi lọc khuyết.

```
> summary(water_potability_no_na)
       pH              Hardness          Solids         Chloramines        Sulfate
 Min.   : 0.2275   Min.   : 73.49   Min.   :  320.9   Min.   : 1.391   Min.   :129.0
 1st Qu.: 6.0897   1st Qu.:176.74   1st Qu.:15615.7   1st Qu.: 6.139   1st Qu.:307.6
 Median : 7.0273   Median :197.19   Median :20933.5   Median : 7.144   Median :332.2
 Mean   : 7.0860   Mean   :195.97   Mean   :21917.4   Mean   : 7.134   Mean   :333.2
 3rd Qu.: 8.0530   3rd Qu.:216.44   3rd Qu.:27182.6   3rd Qu.: 8.110   3rd Qu.:359.3
 Max.   :14.0000   Max.   :317.34   Max.   :56488.7   Max.   :13.127   Max.   :481.0
  Conductivity   Organic_carbon  Trihalomethanes     Turbidity       Potability
 Min.   :201.6   Min.   : 2.20   Min.   :  8.577   Min.   :1.450   Min.   :0.0000
 1st Qu.:366.7   1st Qu.:12.12   1st Qu.: 55.953   1st Qu.:3.443   1st Qu.:0.0000
 Median :423.5   Median :14.32   Median : 66.542   Median :3.968   Median :0.0000
 Mean   :426.5   Mean   :14.36   Mean   : 66.401   Mean   :3.970   Mean   :0.4033
 3rd Qu.:482.4   3rd Qu.:16.68   3rd Qu.: 77.292   3rd Qu.:4.514   3rd Qu.:1.0000
 Max.   :753.3   Max.   :27.01   Max.   :124.000   Max.   :6.495   Max.   :1.0000
```

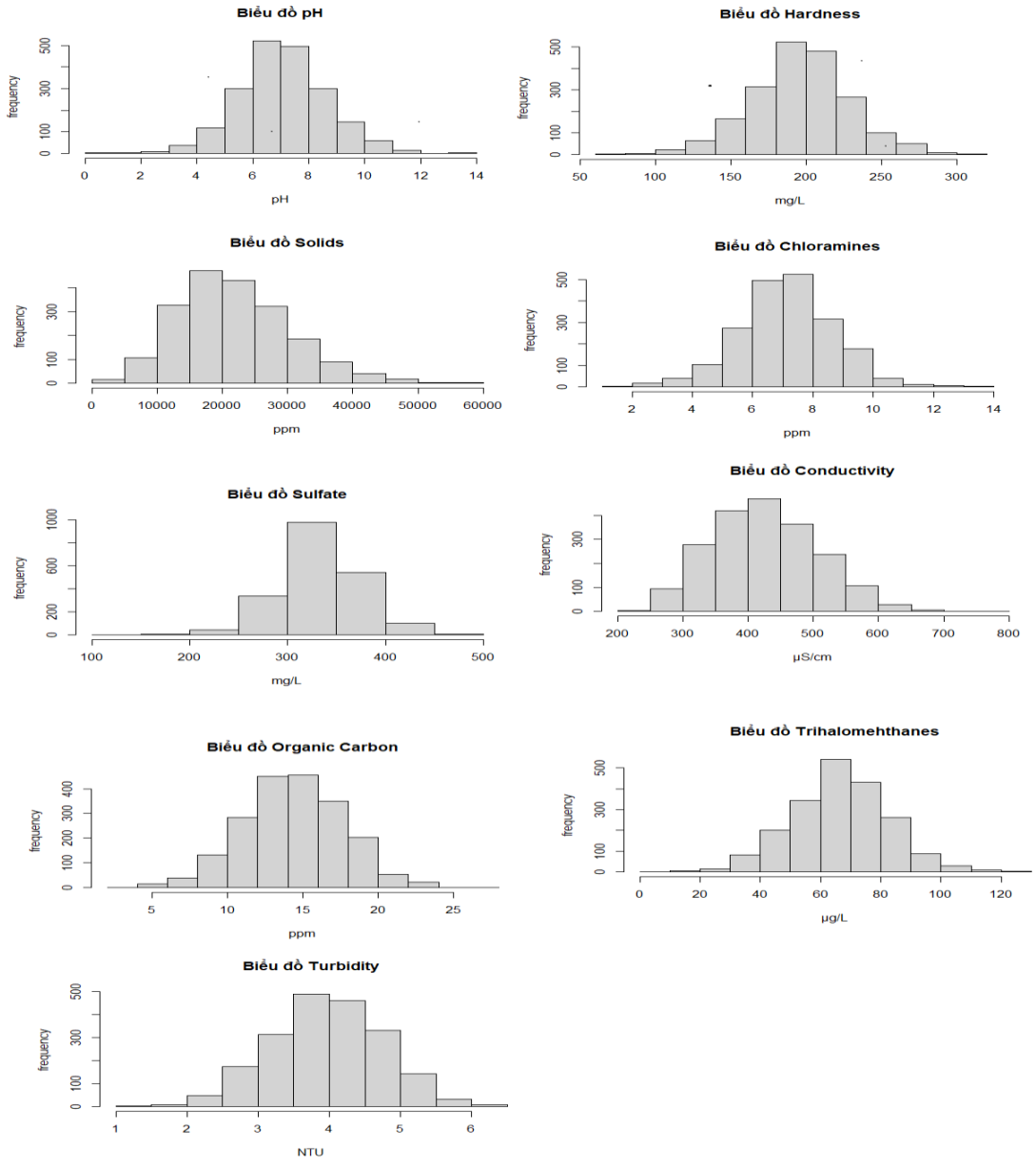
Hình 8: Tổng quan các số liệu thống kê của dữ liệu sau khi lọc khuyết

Tiếp theo, ta dung lệnh as.factor cho biến potability để khai báo đó là biến phụ thuộc để sử cho thống kê mô tả và thống kê suy diễn ở các phần tiếp theo.

IV. THỐNG KÊ MÔ TẢ

1. Histogram

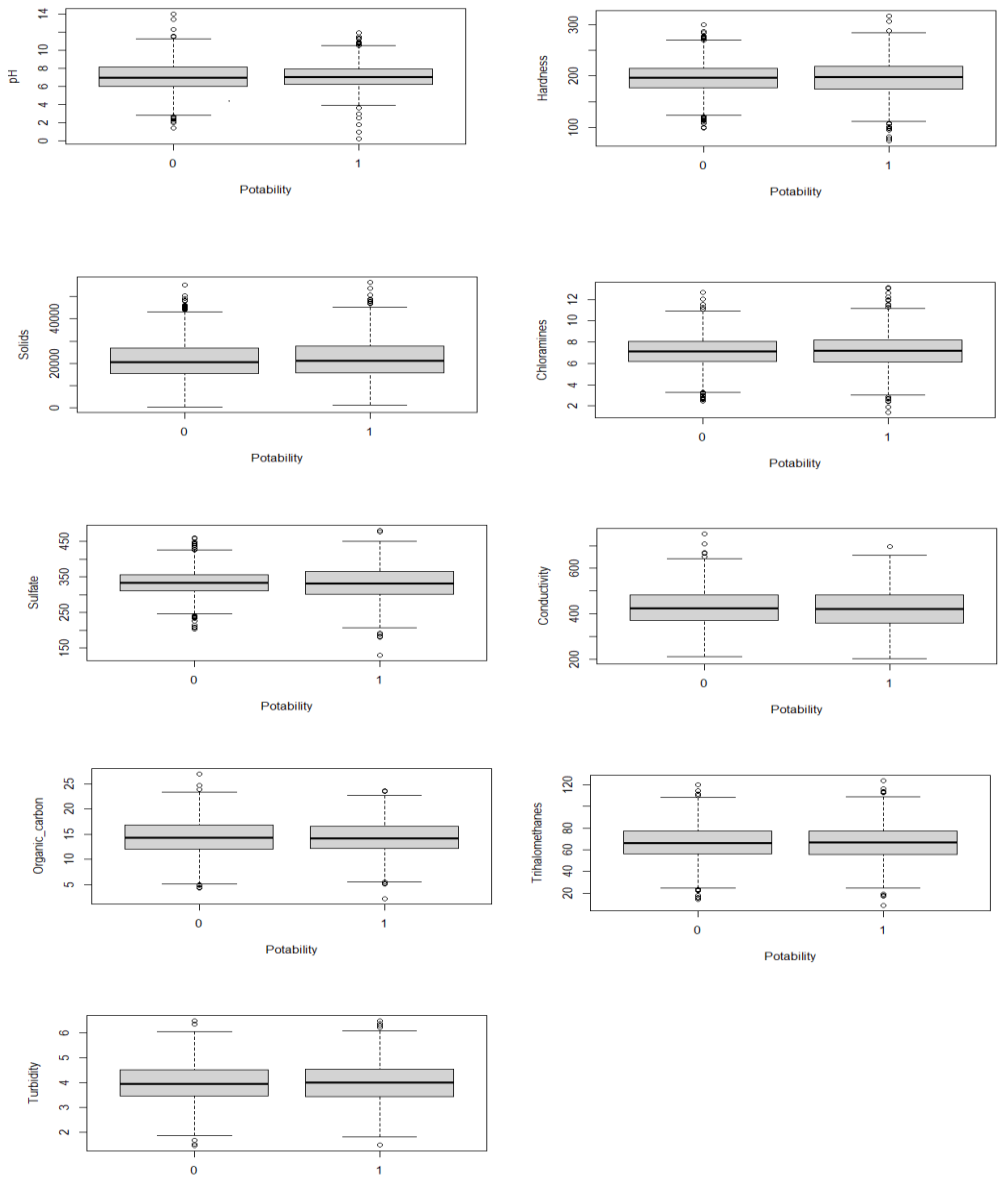
Nhóm đã tiến hành vẽ biểu đồ tần suất (Histogram) nhằm biểu thị mức độ phân phối của bộ dữ liệu được cung cấp.



Hình 9: Biểu đồ cột của 9 biến

Nhận xét: Dựa trên các biểu đồ cột, ta thấy có khá nhiều mẫu nước được thu thập và ghi lại trong dữ liệu thỏa các tiêu chí an toàn cho nước có khả năng sử dụng cho việc uống theo tiêu chuẩn của các cơ quan khoa học quốc tế (với pH nằm trong khoảng 6.5 đến 8.5, độ cứng tối đa 350mg/L). Các yếu tố như Chloramine, Sulfate, Carbon hữu cơ, Trihalomethanes, Solids, Conductivity và Turbidity có số liệu vượt quá mức an toàn của nước uống được không đáng kể.

2. Box plot



Hình 10: Biểu đồ hộp 9 biến chia thành 2 loại nước

Sau khi vẽ xong boxplot, nhóm tóm tắt các số liệu thống kê quan trọng cho từng biểu đồ box plot tương ứng với các biến:

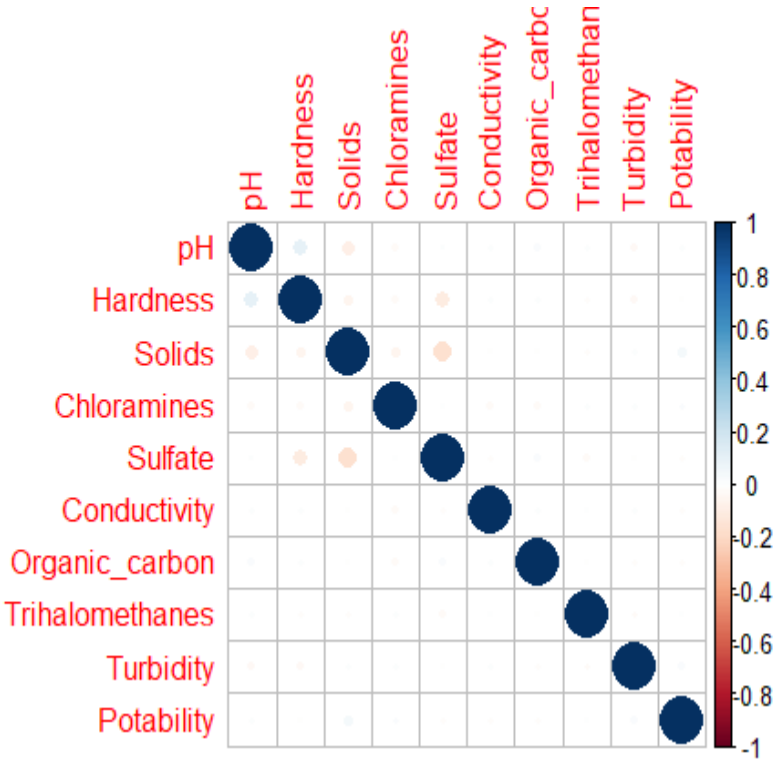
Biến	Kết quả
pH	0 1 Min 1.872573 1.812529 First Quartile 3.444384 3.440564 Median 3.944085 4.007347 Third Quartile 4.498417 4.527463 Maximum 6.064559 6.083772
Hardness	0 1 Min 123.3366 111.4786 First Quartile 177.3073 174.3805 Median 196.7994 197.6175 Third Quartile 214.5676 218.4145 Maximum 270.2095 283.9973
Solids	0 1 Min 320.9426 1198.944 First Quartile 15366.2555 15816.077 Median 20507.3996 21217.159 Third Quartile 26792.4006 27696.134 Maximum 43195.4737 45249.449
Chloramine	0 1 Min 3.318045 3.016033 First Quartile 6.167605 6.106169 Median 7.103718 7.212254 Third Quartile 8.077891 8.181431 Maximum 10.896663 11.170789
Sulfate	0 1 Min 245.7289 206.2472 First Quartile 310.6466 301.7688 Median 332.6156 331.0872 Third Quartile 356.4487 365.6330 Maximum 424.7880 450.9145
Conductivity	0 1 Min 210.3192 201.6197 First Quartile 369.5433 360.2750 Median 424.4795 421.0999 Third Quartile 482.3732 482.2965 Maximum 641.5861 657.5704
Organic_Carbon	0 1 Min 5.15938 5.567693 First Quartile 12.11860 12.148355 Median 14.35183 14.252684 Third Quartile 16.78779 16.561121 Maximum 23.39952 22.641598

Trihalomethanes	0 1 Min 25.05737 24.53277 First Quartile 56.11108 55.75107 Median 66.20612 66.61298 Third Quartile 77.14714 77.37259 Maximum 108.26523 108.84957
Turbidity	0 1 Min 1.872573 1.812529 First Quartile 3.444384 3.440564 Median 3.944085 4.007347 Third Quartile 4.498417 4.527463 Maximum 6.064559 6.083772

Nhận xét: Nhìn chung, ta thấy các biến đều có điểm ngoại lai nhưng chúng đa số đều làm nước không có khả năng sử dụng trong việc uống. Kế đến, các biểu đồ hộp có nhiều giá trị xung quang hoặc bị lệch trên hoặc dưới một ít so với các tiêu chuẩn mà các nghiên cứu khoa học khuyến cáo. Tóm lại, ta có thể thấy rằng, yếu tố uống được của nước phụ thuộc vào hầu hết các yếu tố dù cho một số yếu tố có bị sai khác một phần so với những nghiên cứu đã được đưa ra.

3. Hệ số tương quan

Để thấy mối quan hệ tuyến tính giữa từng biến, chúng ta sẽ vẽ hệ số tương quan của tất cả các biến bằng hàm corrplot.



HÌnh 11: Biểu đồ tương quan của dữ liệu

```
> newwater_potability <- sapply(water_potability_no_na,as.numeric)
> cor_matrix <- cor(newwater_potability)
> cor_matrix
                       pH    Hardness       Solids  Chloramines      Sulfate Conductivity
pH              1.00000000  0.10894811 -0.087614993 -0.024768491  0.010524348  0.014127848
Hardness        0.10894811  1.00000000 -0.053268885 -0.022684975 -0.108520618  0.011730548
Solids         -0.08761499 -0.05326888  1.000000000 -0.051789064 -0.162769204 -0.005197862
Chloramines    -0.02476849 -0.02268498 -0.051789064  1.000000000  0.006254057 -0.028276649
Sulfate         0.01052435 -0.10852062 -0.162769204  0.006254057  1.000000000 -0.016192287
Conductivity    0.01412785  0.01173055 -0.005197862 -0.028276649 -0.016192287  1.000000000
Organic_carbon  0.02837522  0.01322386 -0.005484046 -0.023807630  0.026775563  0.015646727
Trihalomethanes 0.01827788 -0.01540038 -0.015667788  0.014989930 -0.023346904  0.004888475
Turbidity      -0.03584899 -0.03483094  0.019409428  0.013136570 -0.009933881  0.012494892
Potability      0.01453004 -0.00150502  0.040674182  0.020783607 -0.015303149 -0.015495723
                Organic_carbon Trihalomethanes    Turbidity  Potability
pH                  0.028375219     0.018277876 -0.035848994  0.01453004
Hardness            0.013223861    -0.015400382 -0.034830942 -0.00150502
Solids             -0.005484046    -0.015667788  0.019409428  0.04067418
Chloramines        -0.023807630     0.014989930  0.013136570  0.02078361
Sulfate             0.026775563    -0.023346904 -0.009933881 -0.01530315
Conductivity        0.015646727     0.004888475  0.012494892 -0.01549572
Organic_carbon      1.000000000    -0.005667486 -0.015428291 -0.01556703
Trihalomethanes    -0.005667486     1.000000000 -0.020497369  0.00924411
Turbidity          -0.015428291    -0.020497369  1.000000000  0.02268240
Potability         -0.015567030     0.009244110  0.022682396  1.00000000
```

Hình 12: Tóm tắt hệ số tương quan

Nhận xét: Ta nhận thấy rằng mối quan hệ tương quan giữa các biến khá yếu và có thể xem như là không đáng kể. Vì vậy, ta có thể cho rằng các biến gần như độc lập với nhau

V. THỐNG KÊ SUY DIỄN

5.1. Chia lại dữ liệu

Trước khi xây dựng mô hình, ta bắt đầu chia bộ dữ liệu chứa tổng cộng 2011 giá trị quan trắc ra thành hai bộ dữ liệu mới như sau:

- Bộ thứ nhất: chứa 70% dữ liệu, tương ứng với 1408 giá trị quan trắc vào Traindata

- Bộ thứ hai: chứa 30% dữ liệu, tương ứng với 603 giá trị quan trắc vào Testdata

Ngoài ra, ta còn dùng lệnh set.seed để tạo ra các kết quả có thể lặp lại và đảm bảo rằng các giá trị ngẫu nhiên giống nhau được tạo ra mỗi lần chạy chương trình.

5.2. Xây dựng mô hình

Mô hình của chúng ta cũng còn có thể được trình bày dưới dạng:

$$g(\gamma) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i$$

Trong đó:

- $g(\gamma) = \log\left(\frac{\gamma}{1-\gamma}\right)$

- γ: biến phụ thuộc

- X_i: biến độc lập

- β_0: Hệ số chặn (hằng số)

- β_i: Hệ số góc

Nhóm sẽ áp dụng mô hình cho TrainData và hiển thị kết quả.

```
Call:
glm(formula = factor(Potability) ~ pH + Hardness + Solids + Chloramines +
    Sulfate + Conductivity + Organic_carbon + Trihalomethanes +
    Turbidity, family = binomial, data = trainData)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.597e-01  8.935e-01  -1.074    0.283
pH               1.107e-02  3.489e-02   0.317    0.751
Hardness        -2.472e-04  1.709e-03  -0.145    0.885
Solids           2.280e-06  6.477e-06   0.352    0.725
Chloramines      1.491e-02  3.386e-02   0.440    0.660
Sulfate         -3.929e-04  1.343e-03  -0.293    0.770
Conductivity    -3.202e-04  6.801e-04  -0.471    0.638
Organic_carbon   4.492e-03  1.621e-02   0.277    0.782
Trihalomethanes  2.756e-03  3.414e-03   0.807    0.419
Turbidity        1.011e-01  7.042e-02   1.435    0.151

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1899.0  on 1407  degrees of freedom
Residual deviance: 1895.6  on 1398  degrees of freedom
AIC: 1915.6

Number of Fisher Scoring iterations: 4
```

Hình 13: Kết quả của mô hình

5.2. Kiểm tra độ chính xác của mô hình

Sau khi có các hệ số giữa các biến, nhóm sẽ thay các hệ số tương quan vào mô hình và đưa ra dự đoán cho dữ liệu

Bằng cách sử dụng mô hình trên, nhóm giả sử rằng với xác suất lớn hơn 0,5, nước sẽ có khả năng sử dụng được cho việc uống nước. Kết quả sẽ là:

```
> predicted.classes1 <- ifelse(predict(logistic, newdata = trainData, type = "response") >0.5, "0",
"1")
> confusionMatrix(factor(predicted.classes1, levels = c(0,1)), factor(trainData$Potability, levels
= c(0,1)))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0   0   0
         1 840 568

               Accuracy : 0.4034
                 95% CI : (0.3777, 0.4296)
    No Information Rate : 0.5966
    P-Value [Acc > NIR] : 1

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.4034
             Prevalence : 0.5966
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

Hình 14: Kết quả của khả năng dự đoán dựa trên mô hình hồi quy logistic của traindata

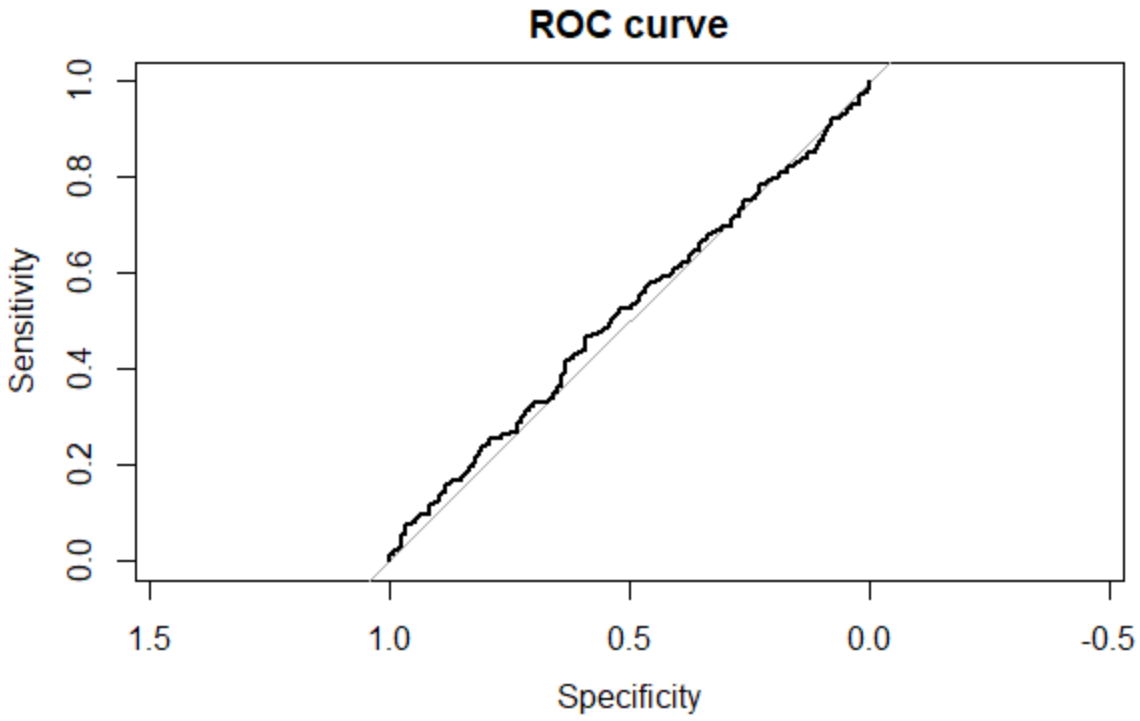
Tương tự ta xây dựng cho testData và thu được kết quả sau đây:

```
> predicted.classes <- ifelse(predict(logistic, newdata = testData, type = "response") >0.5, "0",
"1")
> confusionMatrix(factor(predicted.classes, levels = c(0,1)), factor(testData$Potability, levels =
c(0,1)))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0   0   0
         1 360 243

               Accuracy : 0.403
                 95% CI : (0.3636, 0.4434)
    No Information Rate : 0.597
    P-Value [Acc > NIR] : 1

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.000
            Specificity : 1.000
         Pos Pred Value :   NaN
         Neg Pred Value : 0.403
             Prevalence : 0.597
         Detection Rate : 0.000
   Detection Prevalence : 0.000
      Balanced Accuracy : 0.500

       'Positive' Class : 0
```

Hình 15: Kết quả của khả năng dự đoán dựa trên mô hình hồi quy logistic của Testdata

Nhận xét: ta thấy độ chính xác thấp (chỉ 40%) nhưng ta thấy p-value của mô hình là lớn nhất nên ta thu được kết quả có hiệu suất không quá cao và chỉ ra rằng bộ dữ liệu chưa tương thích lắm với mô hình.

Vì độ chính xác chưa được cao, nhằm cải thiện hiệu suất của mô hình, nhóm sẽ vẽ đường cong ROC để chọn ngưỡng so sánh tốt hơn.



Hình 16: Mô hình đường cong ROC

Lấy ngưỡng so sánh mới từ đường cong ROC (điểm trên cùng bên trái), áp dụng cho mô hình và dự đoán lại. Kết quả thu được cho 2 bộ dữ liệu là:

```
> predicted.classes1 <- ifelse(predict(logistic, newdata = trainData, type = "response") > best_poi
nt, "1", "0")
> confusionMatrix(factor(predicted.classes1,levels = c(0,1)), factor(trainData$Potability, levels=
c(0,1)))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 503 309
         1 337 259

               Accuracy : 0.5412
                 95% CI : (0.5147, 0.5675)
    No Information Rate : 0.5966
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.0544

 Mcnemar's Test P-Value : 0.2881

            Sensitivity : 0.5988
            Specificity : 0.4560
         Pos Pred Value : 0.6195
         Neg Pred Value : 0.4346
             Prevalence : 0.5966
         Detection Rate : 0.3572
   Detection Prevalence : 0.5767
      Balanced Accuracy : 0.5274

       'Positive' Class : 0
```

Hình 17: Độ chính xác của Traindata sau khi thay đổi ngưỡng so sánh

```
> predicted.classes <- ifelse(predict(logistic, newdata = testData, type = "response") > best_poin
t, "1", "0")
> confusionMatrix(factor(predicted.classes,levels = c(0,1)), factor(testData$Potability, levels= c
(0,1)))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 213 129
         1 147 114

               Accuracy : 0.5423
                 95% CI : (0.5016, 0.5826)
    No Information Rate : 0.597
    P-Value [Acc > NIR] : 0.9972

                  Kappa : 0.0601

 Mcnemar's Test P-Value : 0.3062

            Sensitivity : 0.5917
            Specificity : 0.4691
         Pos Pred Value : 0.6228
         Neg Pred Value : 0.4368
             Prevalence : 0.5970
         Detection Rate : 0.3532
   Detection Prevalence : 0.5672
      Balanced Accuracy : 0.5304

       'Positive' Class : 0
```

Hình 18: Độ chính xác của Testdata sau khi thay đổi ngưỡng so sánh

Nhận xét: Sau khi thay đổi ngưỡng so sánh, ta thấy độ chính xác đã tăng đáng kể dù độ chính xác sau cùng vẫn chỉ ở mức trung bình. Ta có thể kết luận rằng việc thay đổi ngưỡng so sánh đã làm cho mô hình phù hợp hơn với bộ dữ liệu.

5.3. Kiểm tra giả định

Chúng ta có thể thấy rằng mặc dù chúng ta chọn ngưỡng từ đường cong ROC nhưng tình hình vẫn chưa đã cải thiện nhiều. Vì vậy, chúng ta sẽ thực hiện kiểm định Hosmer và Lemeshow để xem mức độ phù hợp của mô hình:

```
> fitted_numeric <- as.numeric(as.character(logistic$fitted.values))
> potability_int <- as.integer(as.character(trainData$Potability))
> hl <- hoslem.test(potability_int, fitted_numeric, g=10)
> hl

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  potability_int, fitted_numeric
X-squared = 9.5543, df = 8, p-value = 0.2977
```

Hình 19: Kiểm định Hosmer và Lemeshow cho mô hình

Nhận xét: Như chúng ta đã biết, giá trị p lớn, lớn hơn 0,05, cho thấy chúng ta chưa có đủ bằng chứng để kết luận sự thiếu phù hợp. Vì vậy, chúng ta không đủ cơ sở để bác bỏ giả thuyết không cho rằng mô hình phù hợp dữ liệu tốt dù độ chính xác ta thu được chỉ ở mức trung bình.

5.4. Kết luận

Tóm lại, hồi quy logistic được sử dụng để phân tích khả năng uốn được của các mẫu nước. Trong quá trình trực quan hóa dữ liệu, rõ ràng là dữ liệu thể hiện mức độ cao mức độ biến động. Do đó, việc xác định yếu tố nào hoặc phạm vi tương ứng của chúng, có tác động đáng kể nhất đến các cá nhân được khảo sát. Được trang bị kiến thức này, chúng tôi đã tiến hành sử dụng hồi quy logistic tổng quát mô hình kết hợp tất cả các tính năng dữ liệu. Điều đáng chú ý là mô hình này đã đạt được độ chính xác khá thấp, chỉ khoảng 54.12% trên tập kiểm tra và 54% trên tập huấn luyện. Sau đó, chúng tôi đã tiến hành thử nghiệm HosmerLemeshow để đánh giá mức độ phù hợp của mô hình. Đáng chú ý, giá trị p thu được là đặc biệt cao, gần 0.3, cho thấy mô hình của chúng tôi cần cải thiện hơn nữa.

VI. THẢO LUẬN VÀ MỞ RỘNG

1. Ưu điểm

- Xây dựng được mô hình với độ chính xác tương đối.

- Có thể đánh giá tương đối được khả năng sử dụng của một mẫu nước.

2. Hạn chế

- Việc đánh giá khả năng uống được của nước là vấn đề khá phức tạp do nó phụ thuộc khá nhiều yếu tố nhưng không thể loại bỏ các biến bởi vì khả năng uống được của nước phụ thuộc rất nhiều các chỉ số có trong nó. Vì vậy, có nhiều điểm ngoại lai chúng ta chưa loại bỏ nên có ảnh hưởng một phần đến độ chính xác khi ta áp dụng vào mô hình.

- Độ chính xác chưa cao dù độ phù hợp ở mức khá

VII. NGUỒN DỮ LIỆU VÀ NGUỒN CODE

1. NGUỒN DỮ LIỆU

https://www.kaggle.com/datasets/adityakadiwal/water-potability

2. NGUỒN CODE

https://drive.google.com/drive/u/0/folders/1bfO8IqsWqFmQ0aQoqCSeAPoX38oizKTV

1. Nguyễn Tiến Dũng (chủ biên), Nguyễn Đình Huy, Xác suất – Thống kê &

Phân tích số liệu, 2019.

2. Hồi quy Logistic là gì?

https://aws.amazon.com/vi/what-is/logistic-regression/

3. What is logistic regression?

https://www.ibm.com/topics/logistic-regression

4. Jonathan Bartlett, The Hosmer-Lemeshow goodness of fit test for logistic regression

https://thestatsgeek.com/2014/02/16/the-hosmer-lemeshowgoodness-of-fit-test-for-logistic-regression/

5. Shaun Turney, Pearson Correlation Coefficient (r) — Guide & Examples,

https://www.scribbr.com/statistics/pearson-correlation-coefficient