**ĐẠI HỌC QUỐC GIA**

**TRƯỜNG ĐẠI HỌC BÁCH KHOA TP HỒ CHÍ MINH**

**KHOA KỸ THUẬT HÓA HỌC**

🙞∙∙∙☼∙∙∙🙜



**BÁO CÁO BÀI TẬP LỚN**

**MÔN XÁC SUẤT VÀ THỐNG KÊ**

**NHÓM 17 --- Lớp: L10**

**Giảng viên hướng dẫn:**

**Danh sách nhóm:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Sinh viên thực hiện** | **MSSV** | **Lớp** | **Khoa** |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

*Thành phố Hồ Chí Minh – 2024*

**Mục Lục**

[**1.** **Tổng quan dữ liệu** 3](#_Toc166942824)

[**1.1.** **Ngữ cảnh của dữ liệu** 3](#_Toc166942825)

[**1.2.** **Tổng quan về ngữ liệu** 4](#_Toc166942826)

[**2.** **Kiến thức nền** 5](#_Toc166942827)

[**2.1. Bài toán tìm khoảng tin cậy** 5](#_Toc166942828)

[**2.2. Bài toán kiểm định 2 mẫu** 6](#_Toc166942829)

[***2.2.1. Kiểm định tỷ lệ 2 mẫu*** 6](#_Toc166942830)

[***2.2.2. Kiểm định trung bình 2 mẫu*** 6](#_Toc166942831)

[***2.3. Hồi quy logistic*** 7](#_Toc166942832)

[***2.3.1. Khái niệm*** 7](#_Toc166942833)

[***2.3.2. Mô hình hồi quy logistic*** 7](#_Toc166942834)

[**3.** **Tiền xử lý số liệu** 7](#_Toc166942835)

[***3.1.*** ***Đọc dữ liệu*** 7](#_Toc166942836)

[***3.2.*** ***Làm sạch dữ liệu*** 8](#_Toc166942837)

[***3.3.*** ***Làm rõ dữ liệu*** 10](#_Toc166942838)

[**4.** **Thống kê mô tả** 10](#_Toc166942839)

[**5.** **Thống kê suy diễn** 15](#_Toc166942840)

[**5.1.** **Khoảng tin cậy** 15](#_Toc166942841)

[**5.2.** **Kiểm định giả thuyết** 16](#_Toc166942842)

[**5.3.** **Mô hình hồi quy tuyến tính** 19](#_Toc166942843)

[***5.3.1.*** ***Hồi quy tuyến tính giữa 2 biến đơn lẻ:*** 19](#_Toc166942844)

[**5.3.2.** **Hồi quy tuyến tính cho toàn bộ các biến độc lập:** 22](#_Toc166942845)

[**5.3.3.** **Hồi quy tuyến tính cho từng tổ hợp biến có trị số p mạnh:** 24](#_Toc166942846)

[**5.4.** **Hồi quy Logistic** 28](#_Toc166942847)

[**6.** **Thảo luận và mở rộng** 32](#_Toc166942848)

[**6.1. Hồi quy logistic** 32](#_Toc166942849)

[**6.2. Mô hình Decision tree** 32](#_Toc166942850)

1. **Tổng quan dữ liệu**
   1. **Ngữ cảnh của dữ liệu**

Tình trạng ô nhiễm nguồn nước, thiếu nước sạch trong sinh hoạt hằng ngày đang làm tang nguy cơ mắc các bệnh về đường ruột, bệnh ngoài da và một số căn bệnh khác. Theo ước tính của WHO, cho tới nay có khoảng 130 triệu người đang phải đối mặt với việc dùng nước bị nhiễm arsenic với nồng độ cao hơn nồng độ cho phép là 10 mg/lít. Vì vậy việc tiếp cận nguồn nước sạch là yêu cầu cần thiết để đảm bảo sức khỏe. Ở một số vùng, người ta đã chứng minh được rằng đầu tư vào cấp nước và vệ sinh có thể tác động tích cực đến đời sống con người và đồng thời giảm gánh nặng kinh tế vì chi phí khám sức khỏe khi sử dụng nguồn nước bị ô nhiễm có thể cao hơn chi phí thực hiện các biện pháp phòng ngừa.

* 1. **Tổng quan về ngữ liệu**

Tệp dữ liệu được dùng là tệp “water\_potability.csv” chứa các số liệu chất lượng nước của 3276 vùng nước khác nhau và cho thấy các yếu tố ảnh hưởng đến chất lượng nước.

Tệp dữ liệu đã cho thấy chất lượng nước được xác định bởi 10 yếu tố ảnh hưởng qua bảng tóm tắt sau:

|  |  |  |
| --- | --- | --- |
| Tên biến | Đơn vị | Mô tả |
| pH |  | Độ pH của nước (0 đến 14), phạm vi khuyến cáo từ 6.5 đến 8.5 (nằm trong mức cho phép của WHO) |
| Hardness | mg/L | Độ cứng của nước, liên tục, do ion Ca2+ và Mg2+ tạo kết tủa xà phòng |
| Solids | ppm | Tổng lượng chất rắn hòa tan, liên tục, giới hạn mong muốn của TDS là 500mg/l và 1000mg/l được quy định cho nước uống |
| Chloramines | ppm | Lượng clorua trong nước, liên tục, là chất khử trùng trong nước nên mức an toàn là 4ppm |
| Sulfate | mg/L | Lượng sulfat hòa tan, liên tục, xuất hiện trong khoáng chất, kể cả trong không khí, nước ngầm và trong hầu hết các nguồn cung nước ngọt là khoảng từ 3mg/l đến 30mg/l |
| Conductivity | µS/cm | Độ dẫn điện của nước, liên tục, theo WHO giá trị EC không nên vượt quá 400µS/cm |
| Organic cacbon | ppm | Lượng cacbon hữu cơ, liên tục và theo US EPA giá trị TOC trong nước đã qua xử lý/nước uống không nên dưới 2mg/l |
| Trihalomethanes | µg/L | Lượng hợp chất halogen trong nước đã được xử lý, liên tục và nồng độ lên đến 80ppm được coi là an toàn trong nước uống |
| Turbidity | NTU | Độ đục của nước, liên tục và độ đục khuyển nghị của WHO là 5,00 NTU |
| Potability |  | Khả năng uống được của nước (x=0 là không uống được, x=1 là uống được) |

1. **Kiến thức nền**

**2.1.** **Bài toán tìm khoảng tin cậy**

Cho tham số của tổng thể và là các quan sát ngẫu nhiên. Ta gọi khoảng (c, d) là khoảng ước lượng (hay khoảng tin cậy) của tham số với độ tin cậy nếu:

Có thể nói, độ tin cậy cho khoảng ước lượng của tham số chính là xác suất để ta đúng khi ước lượng tham số bằng khoảng (c, d). Ngược lại, xác suất mà ta cho phép sai khi ước lượng được gọi là mức ý nghĩa. Ký hiệu là . Ta có .

***2.1.1. Ước lượng bằng khoảng tin cậy***

Ước lượng bằng khoảng tin cậy là tìm ra khoảng ước lượng cho tham số trong tổng thể sao cho ứng với độ tin cậy (confidence) bằng cho trước,

Phương pháp ước lượng bằng khoảng tin cậy có ưu thế hơn phương pháp ước lượng điểm vì nó làm tăng độ chính xác của ước lượng và còn đánh giá được mức độ tin cậy của ước lượng.

Khả năng mắc sai lầm của phương pháp là .

Trong thống kê, người ta dùng ký hiệu để biểu diễn giá trị xác suất nhỏ. Vì vậy, ký hiệu còn được dùng để biểu diễn độ tin cậy của khoảng ước lượng.

***2.1.2. Tìm khoảng tin cậy cho tham số - kiểm định 1 mẫu***

Phương pháp tìm khoảng tin cậy cho tham số với độ tin cậy

cho trước:

Trước tiên, tìm hàm ước lượng sao cho quy luật phân phối xác suất của G hoàn toàn xác định, không phụ thuộc vào các đối số. Chọn cặp số sao cho và tìm mà suy ra . Biến đổi để tìm được các giá trị sao cho Khi đó khoảng chính là một trong các khoảng tin cậy (confidence interval) cần tìm.

Theo nguyên lý xác suất lớn thì với độ tin cậy đủ lớn, hầu như chắc chắn biến cố sẽ xảy ra trong một phép thử. Vì vậy trong thực tế chỉ cần thực hiện phép thử để có được một mẫu cụ thể

rồi tính giá trị của và ứng với mẫu đã cho sẽ cho ta một khoảng ước lượng thỏa yêu cầu.

Có vô số khoảng ước lượng cho giá trị p của tổng thể theo cách chọn . Đối với bài toán ước lượng tỉ lệ hay ước lượng trung bình thì khoảng ước lượng đối xừng được thiết lập khi chọn

chính là khoảng ước lượng có độ dài ngắn nhất.

Đối với một mẫu đã xác định, khoảng ước lượng đối xứng có độ dài càng hẹp thì độ tin cậy càng thấp. Nếu chúng ta muốn có được ngưỡng sai số nhỏ (khoảng tin cậy hẹp) và độ tin cậy như mong muốn thì chúng ta phải tăng kích thước mẫu hợp lý.

## **2.2.** **Bài toán kiểm định 2 mẫu**

Kiểm định giả thuyết thống kê trong mô hình hồi quy tuyến tính bội.

Bài kiểm định cho ý nghĩa của mô hình hồi quy là một bài kiểm định nhằm xác định xem có hay không một mối quan hệ tuyến tính tồn tại giẵ biến Y và một nhóm biến hồi quy . Khi đó giả thuyết thống kê hợp lý là:

với ít nhất một j

Bác bỏ đồng nghĩa với việc ta chấp nhận có ít nhất một trong các biến hồi quy có ảnh hưởng đến kể đến mô hình.

Mục đích của kiểm định để xem 2 mẫu có sự khác biệt gì, có thể kiểm định một phía hay hai phía. Các dạng kiểm định 2 mẫu thường gặp: kiểm định tỷ lệ, kiểm định trung bình.

### ***2.2.1.*** ***Kiểm định tỷ lệ 2 mẫu***

Cho tổng thể X có tỉ lệ các phần tử có đặc điểm A là  và tổng thể Y có tỷ lệ các phần tử có đặ điểm A là , so sánh tỷ lệ 2 tổng thể này với ý nghĩa

Phương pháp:

Bước 1: xác định giả thuyết và

Bước 2: sử dụng các công thức phù hợp để tìm miền bác bỏ, tiêu chuẩn kiểm định

Bước 3: kết luận

- Nếu thì ta có thể bác bỏ giả thuyết , chấp nhận giả thuyết

- Nếu thì ta không bác bỏ ý kiến (chưa có đủ bằng chứng để bác bỏ )

### ***2.2.2. Kiểm định trung bình 2 mẫu***

Cho tổng thể X có trung bình là là tổng thể Y có trung bình là , so sánh tỷ lệ 2 tổng thể này với ý nghĩa , với 2 mẫu độc lập

Phương pháp: tương tự như với bài toán kiểm định tỷ lệ 2 mẫu điểm khác là cần sử dụng các công thức phù hợp để tìm miền bác bỏ và tiêu chuẩn kiểm định.

## **2.3. Hồi quy logistic**

### ***2.3.1. Khái niệm***

Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa 2 yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn như có hoặc không.

### ***2.3.2. Mô hình hồi quy logistic***

Hàm hồi quy logistic: hồi quy logistic là một mô hình thống kê sử dụng hàm logistic, hay hàm logit trong toán học làm phương trình giữa x và y. Hàm logit ánh xạ y làm hàm sigmoid của x:

Như bạn có thể thấy, hàm logit chỉ trả về các giá trị giữa 0 và 1 cho biến phụ thuộc, dù giá trị của biến độc lập là gì. Đây là cách hồi quy logistic ước tính giá trị của biến phụ thuộc. Phương pháp hồi quy logistic cũng lập mô hình phương trình giữa nhiều biến độc lập và một biến phụ thuộc.

Có thể sử dụng một giá trị ngưỡng t (0 < t < 1)

Nếu : dự báo y = 1

Nếu : dự báo y = 0

Thông thường chọn t = 0.5

Phương pháp hồi quy logistic cũng lập mô hình phương trình giữa nhiều biến độc lập và một biến phụ thuộc. Trong nhiều trường hợp, nhiều biến giải thích ảnh hưởng đến giá trị của một biến phụ thuộc. Để lập mô hình các tập dữ liệu đầu vào như vậy, công thức hồi quy logistic phải giả định mối quan hệ tuyến tính giữa các biến độc lập khác nhau. Có thể sửa đổi hàm sigmoid và tính toán biến đầu ra cuối cùng như sau:

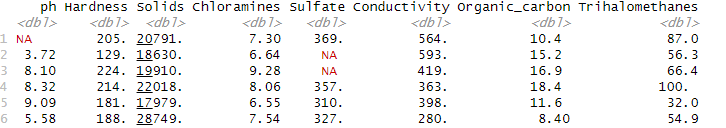
Ký hiệu đại diện cho hệ số hồi quy. Mô hình logit có thể đảo ngược tính toán các giá trị hệ số này khi bạn cho nó một tập dữ liệu thực nghiệm đủ lớn có các giá trị đã xác định của cả hai biến phụ thuộc và biến độc lập.

1. **Tiền xử lý số liệu**
   1. ***Đọc dữ liệu***

Đọc dữ liệu: **“ water\_potibility.csv”**

Hình 1: Code R và kết quả khi đọc dữ liệu và xem 6 dòng đầu tiên của dữ liệu

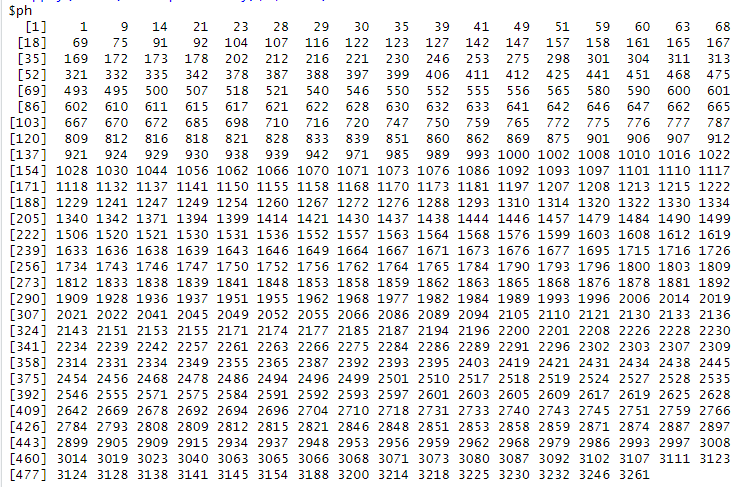


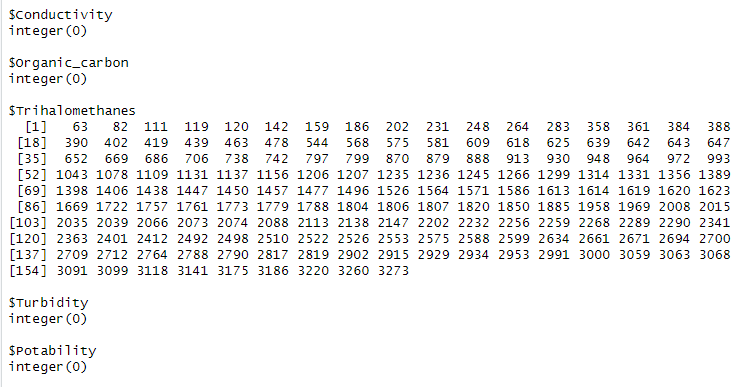


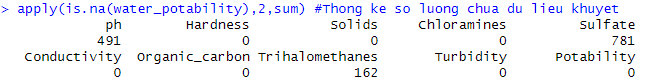
* 1. ***Làm sạch dữ liệu***

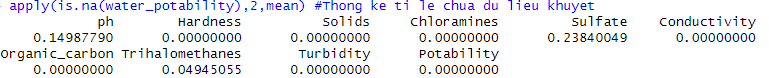
Hình 2: Code R và kết quả khi kiểm tra dữ liệu khuyết





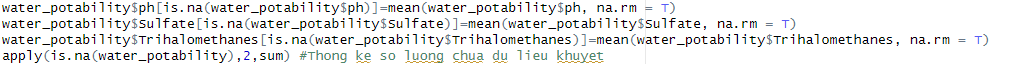


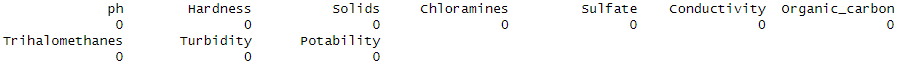




Nhận xét: Dựa vào kết quả thu được khi kiểm tra dữ liệu khuyết, ta nhận thấy có 491 dữ liệu khuyết tại biến ph, 781 dữ liệu khuyết tại biến Sulfate và 162 dữ liệu khuyết tại biến Trihalomethanes. Vậy nên ta cần xử lý các dữ liệu khuyết đó.

Hình 3: Code R khi thay thế giá trị trung bình ở các quan sát còn lại của biến ph, Sulfate,Trihalomethanes tại vị trí chứa dữ liệu khuyết và kiểm tra lại xem còn dữ liệu khuyết hay không

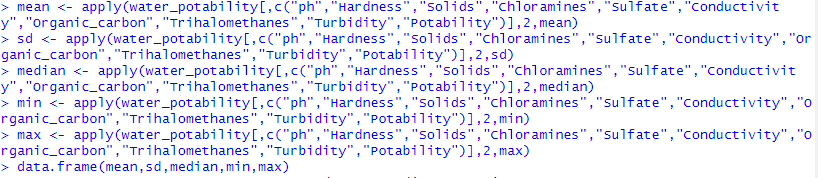


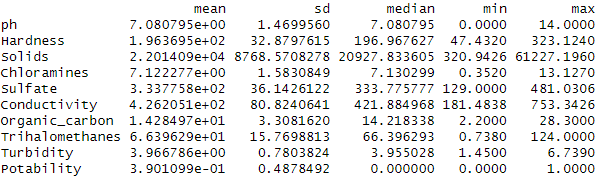


Nhận xét: Phương pháp dùng để xử lí dữ liệu khuyết được đề xuất là thay thế giá trị trung bình ở các quan sát còn lại của các biến ph, Sulfate, Trihalomethanes tại vị trí chứa dữ liệu khuyết và sau khi xử lí thì không còn dữ liệu khuyết.Với tỉ lệ dữ liệu khuyết vượt quá 10% nên ta không nên dùng lệnh ” na.omit “ để loại bỏ dữ liệu khuyết.

* 1. ***Làm rõ dữ liệu***

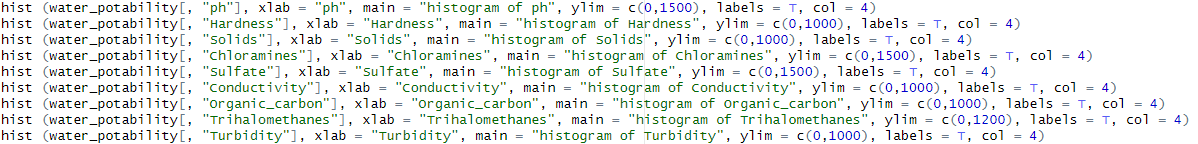
Hình 4: Code R và kết quả tính các giá trị thống kê mô tả (trung bình, độ lệch chuẩn, trung vị, min, max)

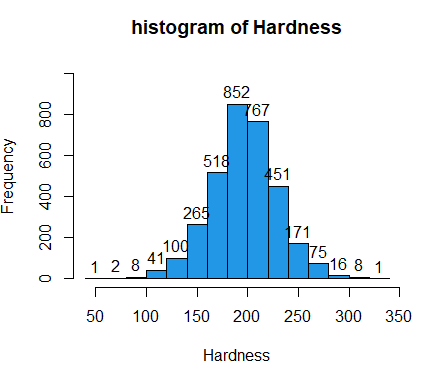
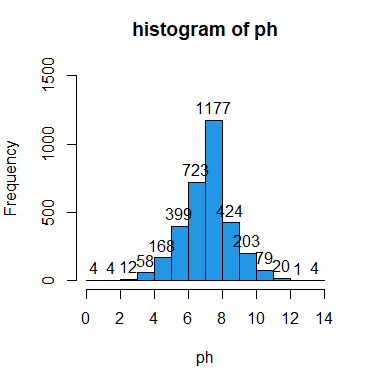


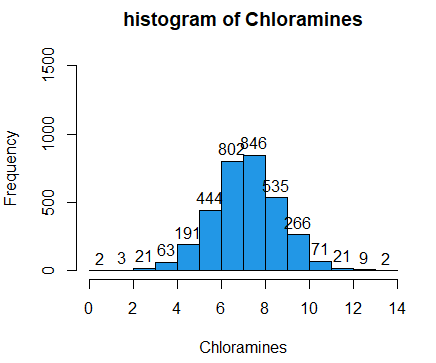
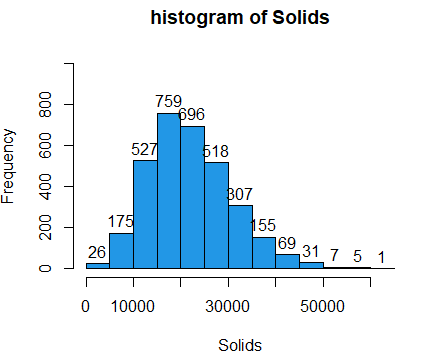


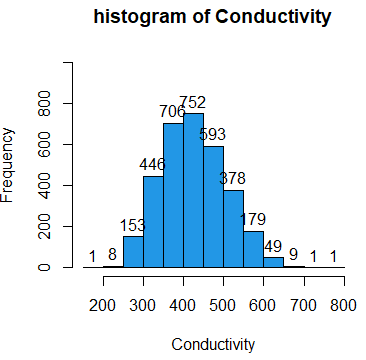
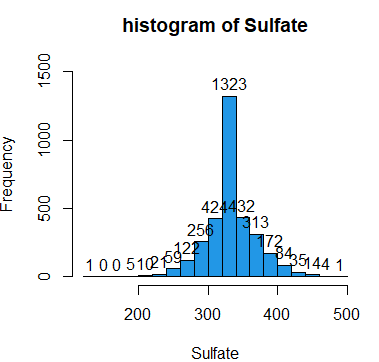
1. **Thống kê mô tả**

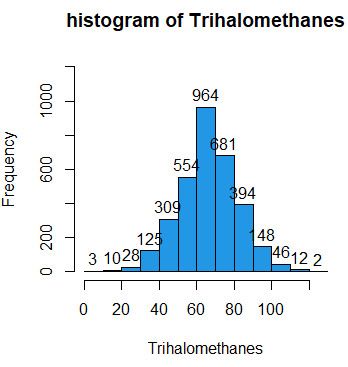
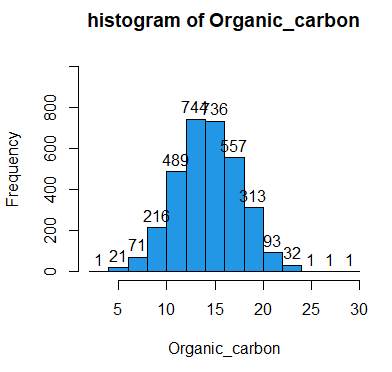
Hình 5: Code R và kết quả khi vẽ biểu đồ histogram thể hiện phân phối của các biến

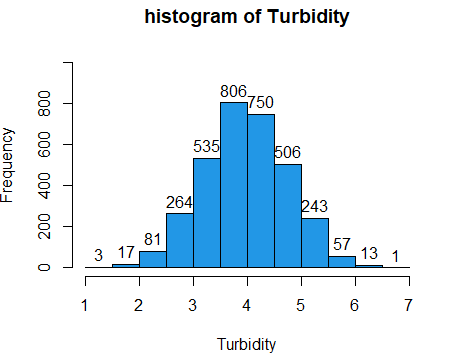






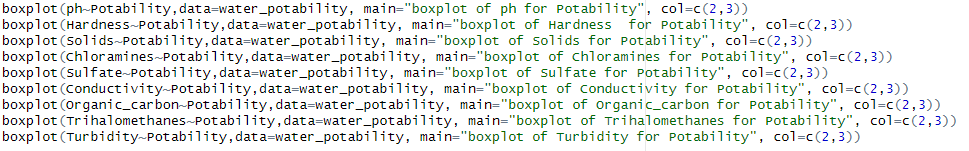


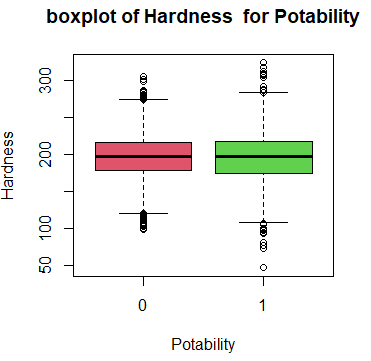
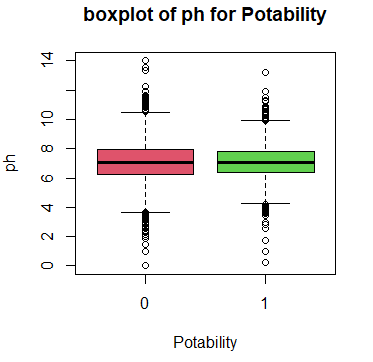


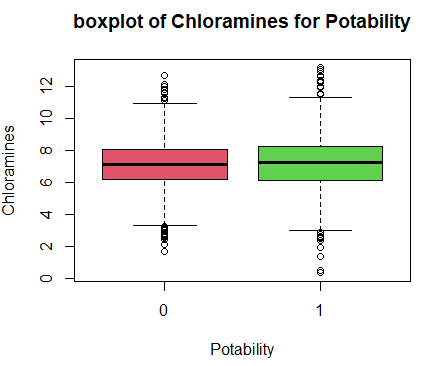
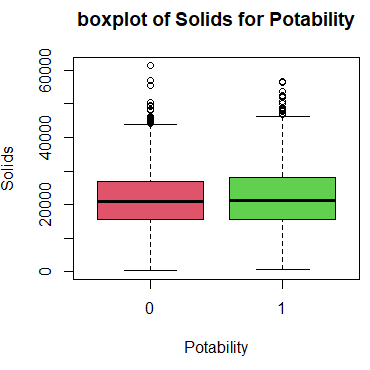


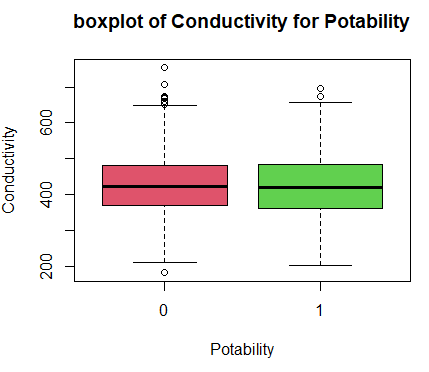
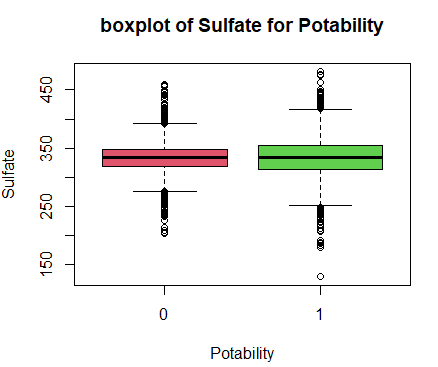
Nhận xét: Ta thấy đồ thị histogram của các biến ph, Hardness, Chloramines, Sulfate, Conductivity, Organic\_carbon, Trihalomethanes, Turbidity tuân theo phân phối chuẩn. Trong đó đồ thị histogram của biến Solids có xu hướng hơi lệch phải, cho thấy tổng chất rắn hòa tan gần bằng nhau , chỉ một số ít có tổng chất rắn hoà tan lớn hơn.

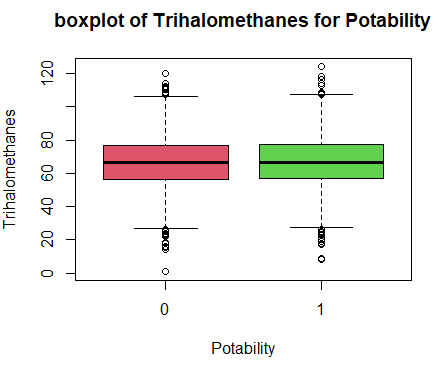
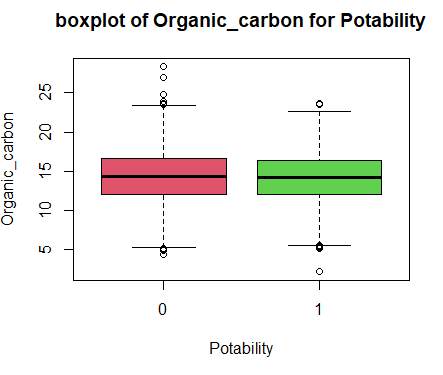
Hình 6: Code R và kết quả khi vẽ biểu đồ boxplot

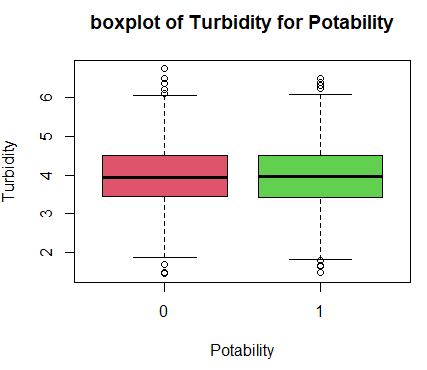












Nhận xét: Đồ thị boxplot biểu diễn phân bố của dữ liệu. Nó cho biết vị trí của các phần tư vị (25%, 50%, và 75%), trung vị, phạm vi tương đối của dữ liệu: chiều dài của hộp (khoảng giữa Q1 và Q3) , giá trị ngoại lai thường được biểu diễn dưới dạng các điểm nằm ngoài phạm vi, phân phối của dữ liệu như nó có lệch phải, lệch trái hay đối xứng.

Ví dụ độ pH ảnh hướng đến khả năng uống được:

*Với nước không uống được*: *Với nước uống được:*

Max= 14 Max=13

Min= 0 Min=0.5

Median= 7 Median=7

G1= 6 G1=6.5

G3= 8 G3=7.5

1. **Thống kê suy diễn**
   1. **Khoảng tin cậy**

***Mục tiêu:***

Ta tiến hành xét khoảng tin cậy đối với từng cột dữ liệu biến độc lập (ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic\_carbon, Trihalomethanes và Turbidity) với mức tin cậy là 95% và xem xét số lượng các điểm dữ liệu nằm trong, nằm ngoài khoảng tin cậy để dự đoán sự biến động dữ liệu trong các cột dữ liệu.

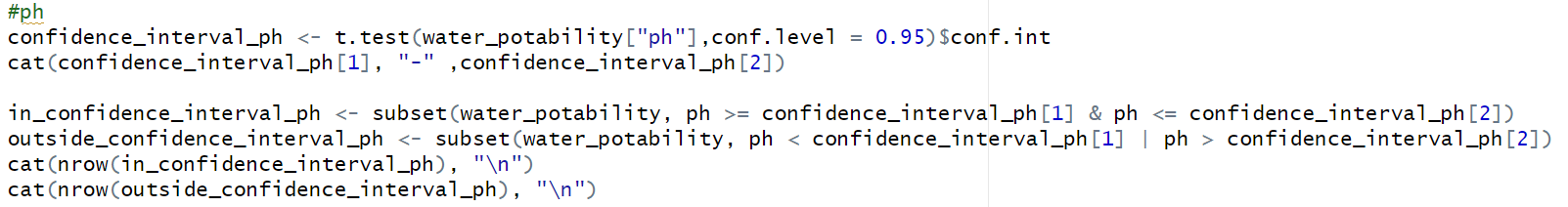
***Ý tưởng:***

Sử dụng lệnh t.test với hai input đầu vào là một tập dữ liệu cần xét và conf.level (mức tin cậy là 95%) và giá trị trả về bao gồm cận trên và cận dưới của khoảng tin cậy.

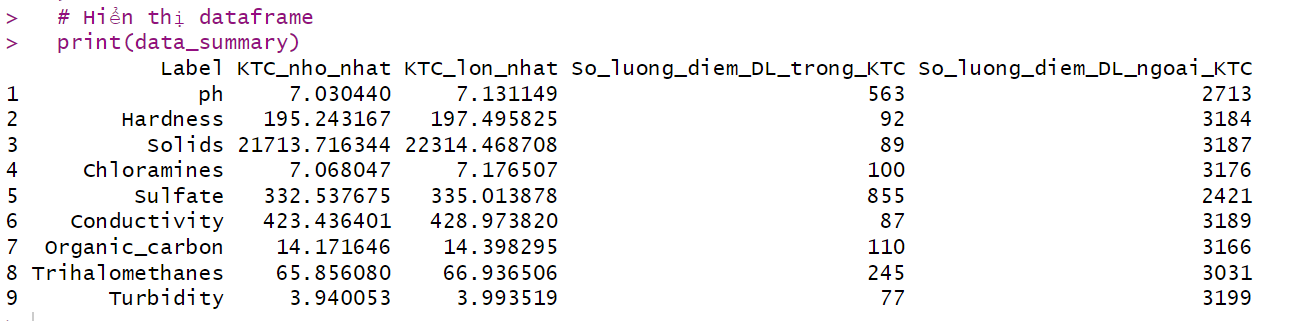
Sử dụng subset để lọc ra điểm nằm trong khoảng tin cậy và ngoài khoảng tin cậy. Sử dụng nrow để tính số lượng điểm dữ liệu trong các khoảng ấy.

***Hiện thực và kết quả:***

Hình 7: Đoạn code về khoảng tin cậy cho biến pH



Các cột dữ liệu còn lại ta làm tương tự

Hình 8: Kết quả có được khoảng tin cậy cho từng cột dữ liệu

**Nhận xét:**

Qua bảng tổng hợp về Khoảng tin cậy của các cột dữ liệu với mức độ tin cậy là 95%, ta nhận thấy các khoảng tin cậy đều có phạm vi rất nhỏ so với toàn bộ phạm vi của các cột dữ liệu (ví dụ: PH có phạm vi từ 7.0304 đến 7.1311; Hardness có phạm vi từ 195.2432 đến 197.4958)

Số lượng điểm dữ liệu trong khoảng tin cậy là rất nhỏ so với số lượng điểm dữ liệu nằm ngoài khoảng tin cậy. Khi xét tỷ lệ số lượng điểm dữ liệu trong Khoảng tin cậy và Tỷ lệ số lượng điểm dữ liệu nằm ngoài Khoảng tin cậy, ta nhận thấy tỷ lệ của Sulfate có tỷ lệ lớn nhất (855/2421 ≈ 0.3532) và nhỏ nhất là tỷ lệ của Turbidity (77/3199 ≈ 0.0241).

**Kết luận:**

Qua việc phân tích khoảng tin cậy của các cột dữ liệu, ta thấy rằng tập dữ liệu ở các cột dữ liệu đều có tỷ lệ điểm dữ liệu nằm trong Khoảng tin cậy, điều này thể hiện rằng tập dữ liệu trong các cột dữ liệu đều có sự biến động đáng kể và có thể ảnh hưởng đến kết quả của các phép phân tích tính toán sau này.

* 1. **Kiểm định giả thuyết**

**Mục tiêu:**

Ta tiến hành kiểm định giả thuyết để so sánh trung bình hai mẫu giữa hai nhóm khác nhau trong một tập dữ liệu với mức ý nghĩa 0.05. Cụ thể ta sẽ kiểm định về sự khác biệt giữa một cột dữ liệu độc lập theo cột dữ liệu mục tiêu (Potability) khi giá trị điểm dữ liệu trong cột dữ liệu mục tiêu là 0 (không thể uống được) hay 1 (có thể uống được). Vì vậy, ta có:

Giả thuyết (H0): μ0 = μ1 (Không có sự khác biệt đáng kể về trung bình trong cột dữ liệu độc lập giữa nước uống có thể uống được và nước uống không thể uống được.)

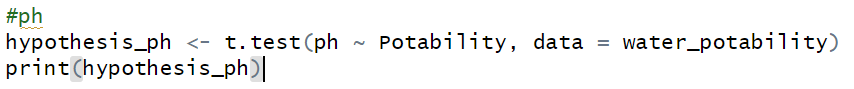
Đối thuyết (H1): μ0 ≠ μ1 (Có sự khác biệt đáng kể về trung bình trong cột dữ liệu độc lập giữa nước uống có thể uống được và nước uống không thể uống được.)

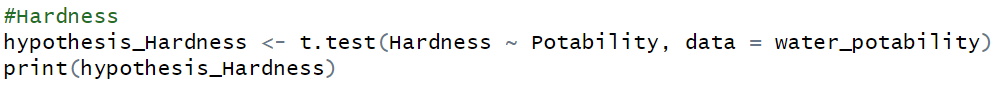
Với cột dữ liệu độc lập sẽ là các cột: ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic\_carbon, Trihalomethanes và Turbidity.

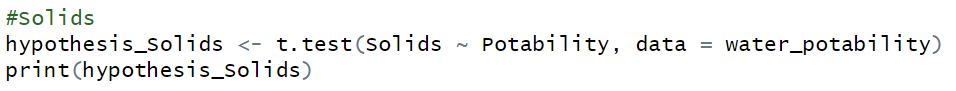
**Ý tưởng:**

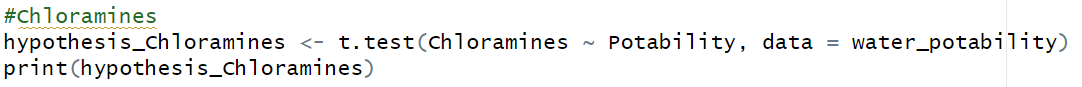
Sử dụng t.test với input đầu vào là một cột dữ liệu biến độc lập và tập dữ liệu.

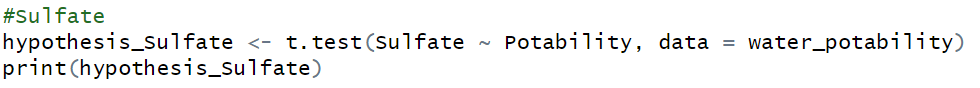
**Hiện thực và kết quả:**

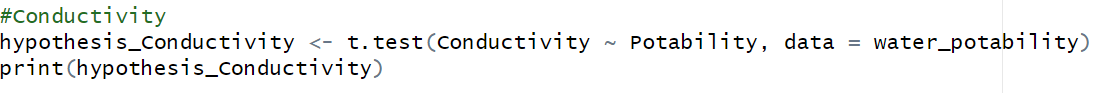
Hình 9: Đoạn code kiểm định cho từng cột dữ liệu biến độc lập.

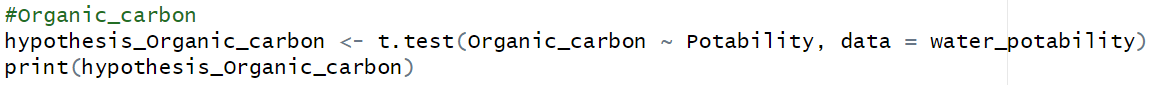


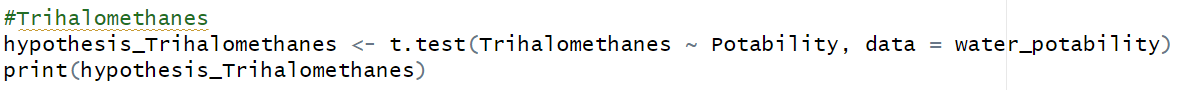


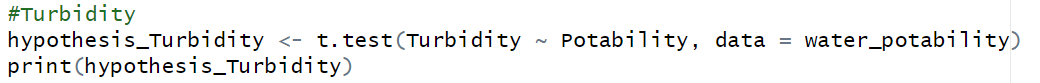






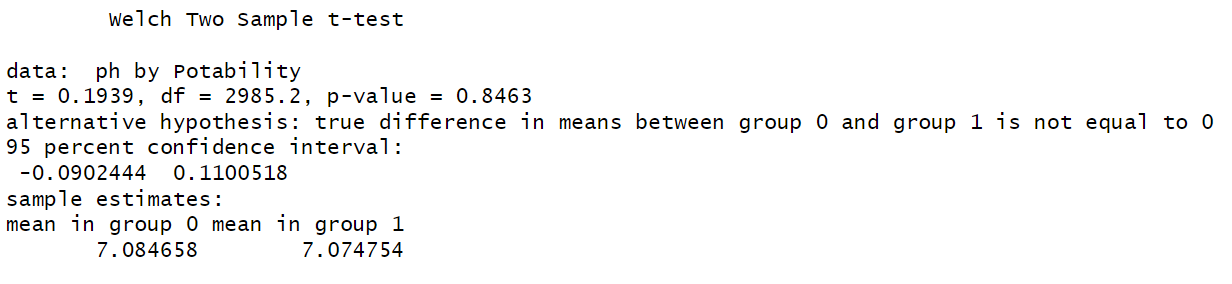




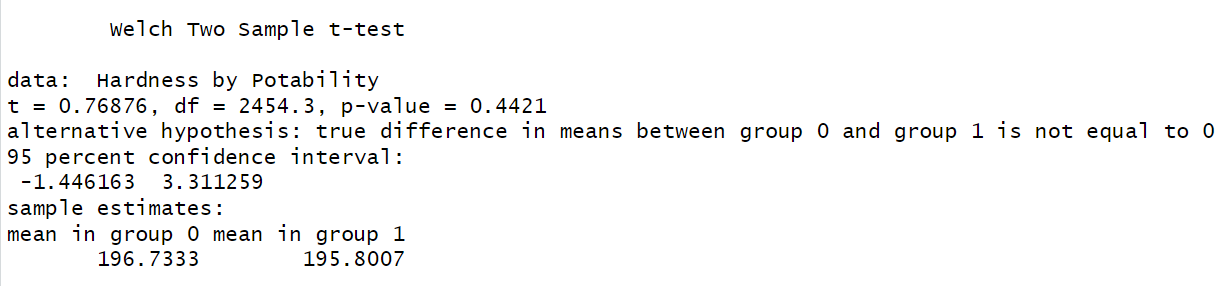


Hình 10 : kết quả kiểm định cho từng cột dữ liệu biến độc lập

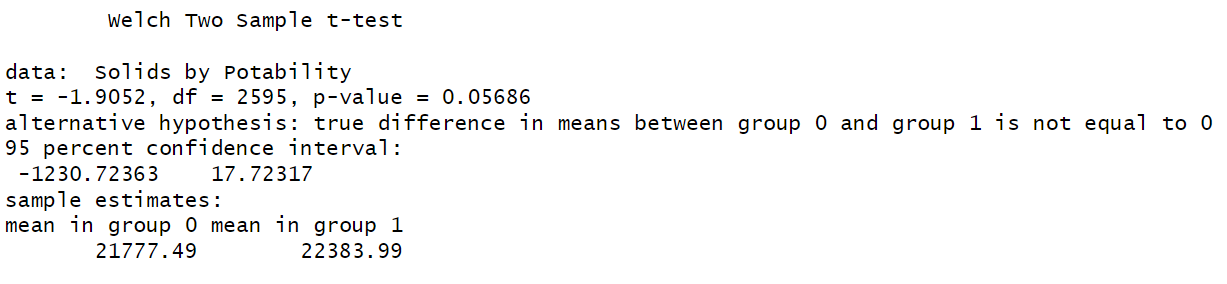
pH



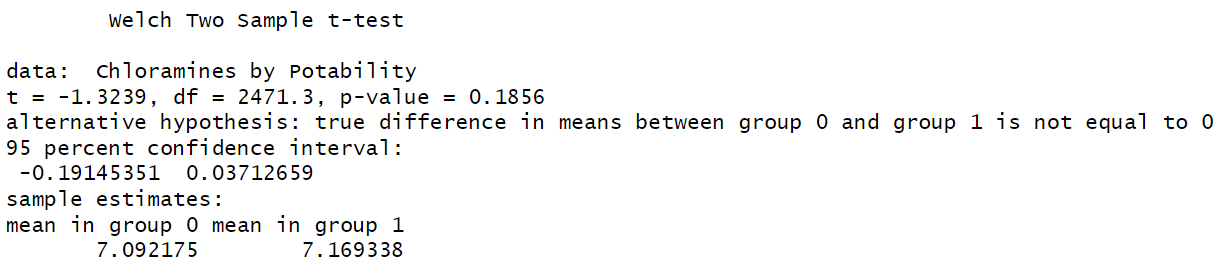
Hardness:



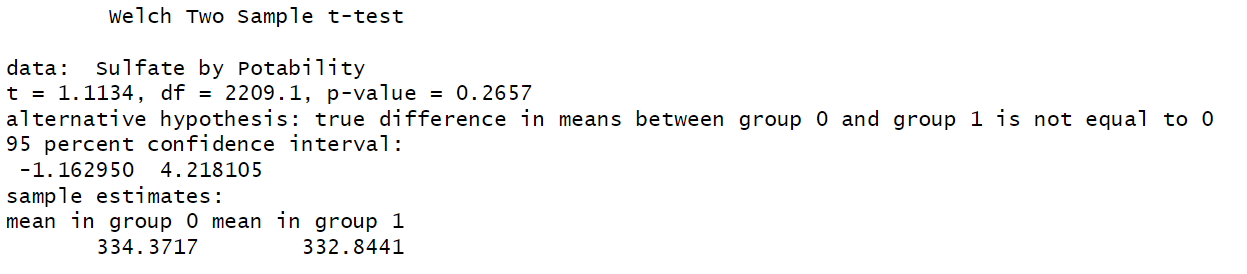
Solids:



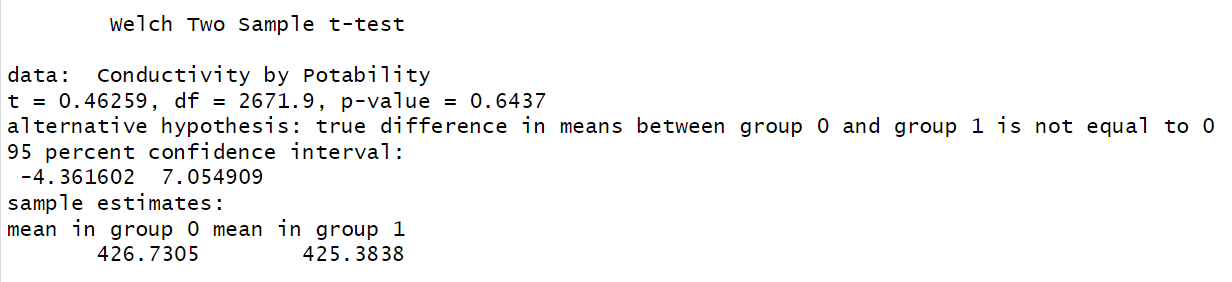
Chloramines:



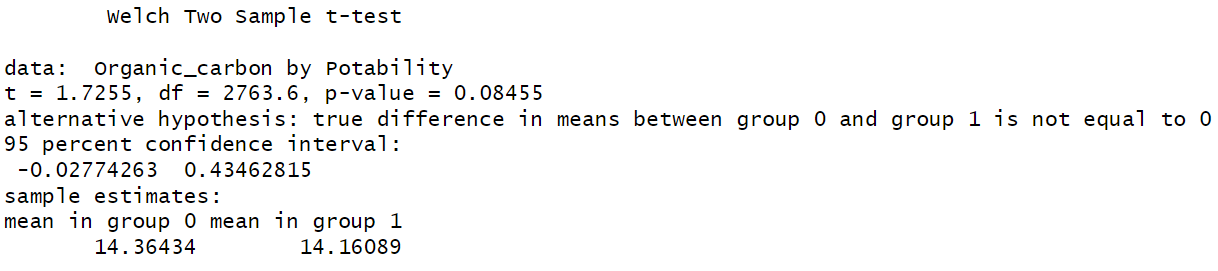
Sulfate:



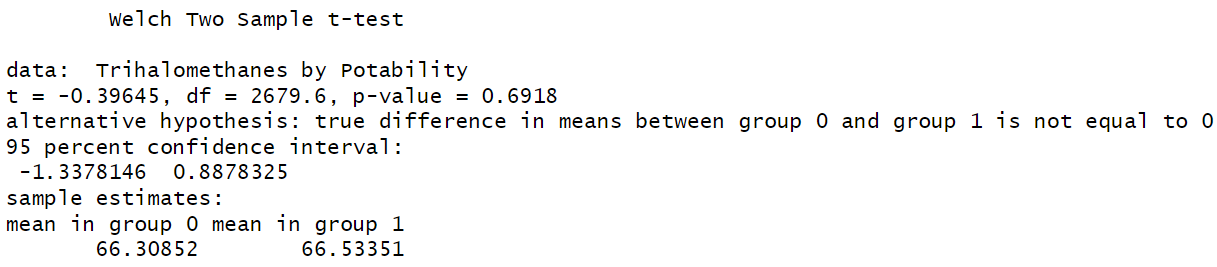
Conductivity:



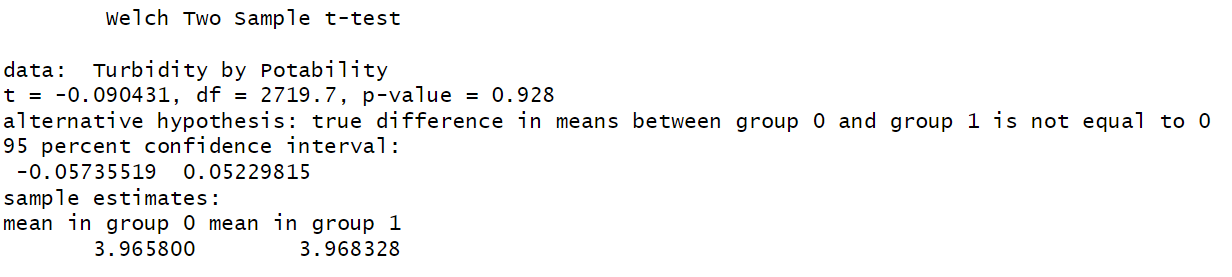
Organic\_carbon:



Trihalomethanes:



Turbidity:



**Nhận xét:**

Tất cả các tập dữ liệu đều sử dụng phương pháp Welch t.test (dùng để so sánh giá trị trung bình giữa hai tập dữ liệu khi chúng có phương sai không đồng nhất).

Ta nhận thấy ở tất cả các cột dữ liệu độc lập, ta nhận thấy rằng p-value của các cột dữ liệu độc lập đều lớn hơn 0.05 (cột dữ liệu có p-value gần nhất là cột Solids với p-value = 0.05686).

Qua việc quan sát các giá trị trung bình của các cột dữ liệu độc lập, ta nhận thấy các giá trị này ở các cột đều có giá trị tương đối gần nhau.

**Kết luận:**

Qua việc thực hiện kiểm định 2 mẫu với từng giá trị 0 và 1 của cột Potability, ta nhận thấy rằng ở tất cả các cột dữ liệu độc lập đều lớn hơn 0.05 cho thấy rằng không có đủ bằng chứng để bác bỏ giả thuyết H0 trong mọi trường hợp. Tức là ta thừa nhận với mức tin cậy 95% rằng, không có sự khác biệt đáng kể giữa giá trị trung bình của các điểm dữ liệu của các cột độc lập khi điểm dữ liệu Potability là 0 (không thể uống được) và là 1 (có thể uống được). Điều này cũng đã được kiểm nghiệm một lần nữa khi quan sát giá trị trung bình giữa hai tập dữ liệu này.

* 1. **Mô hình hồi quy tuyến tính**
     1. ***Hồi quy tuyến tính giữa 2 biến đơn lẻ:***

**Mục tiêu:**

Ta sẽ tiến hành kiểm tra mức độ tương quan tuyến tính giữa các biến độc lập với nhau và giữa biến phụ thuộc (Potability) và một biến độc lập.

**Ý tưởng:**

Ta sẽ sử dụng các hàm trong hai thư viện dplyr và corrplot.

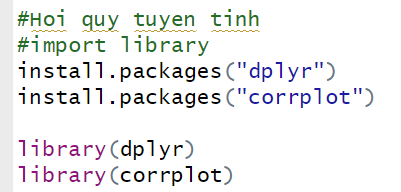
Sử dụng hàm Cor với đầu vào là tập dữ liệu gốc, biến phụ thuộc Potability và tập các biến độc lập.

Sử dụng Corrplot để tạo biểu đồ ma trận tương quan dưới dạng các hình tròn.

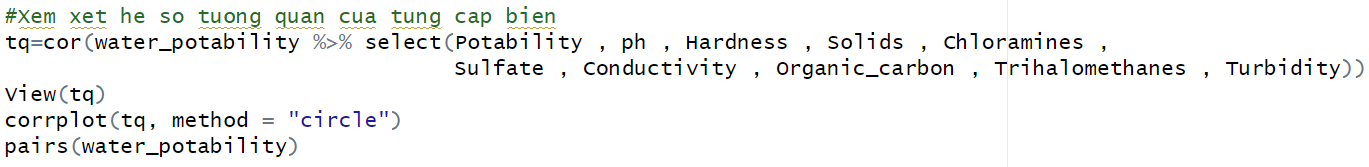
Sử dụng lệnh Pairs để tạo biểu đồ ma trận tương quan giữa các cặp biến bằng các điểm dữ liệu.

**Hiện thực và kết quả:**

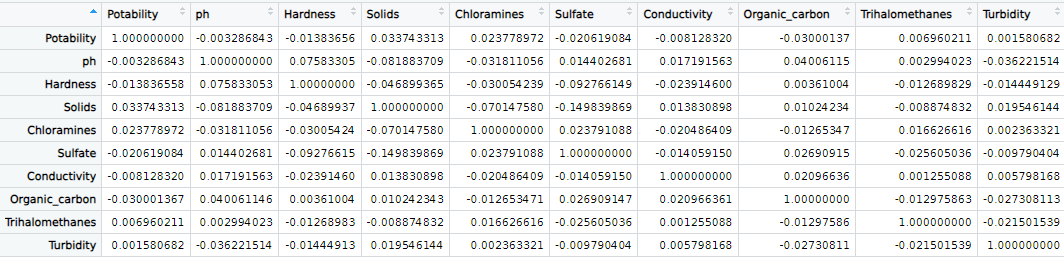
Ta sẽ sử dụng các thư viện sau:

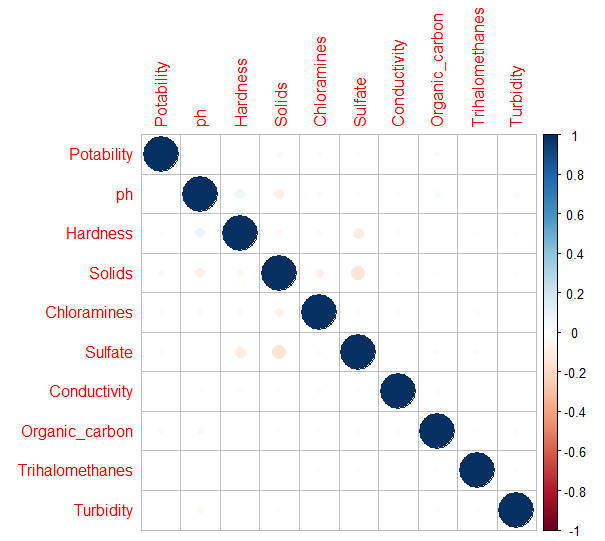


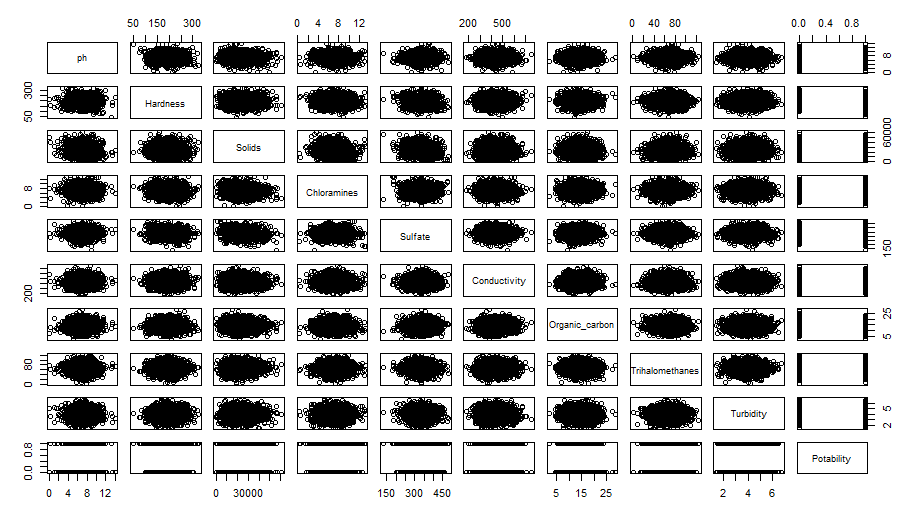
Hình 11: Xem xét hệ số tương quan của từng cặp biến:

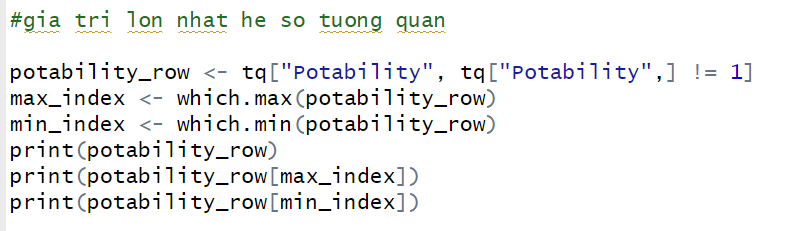


Hình 12: Kết quả hệ số tương quan của từng cặp biến:

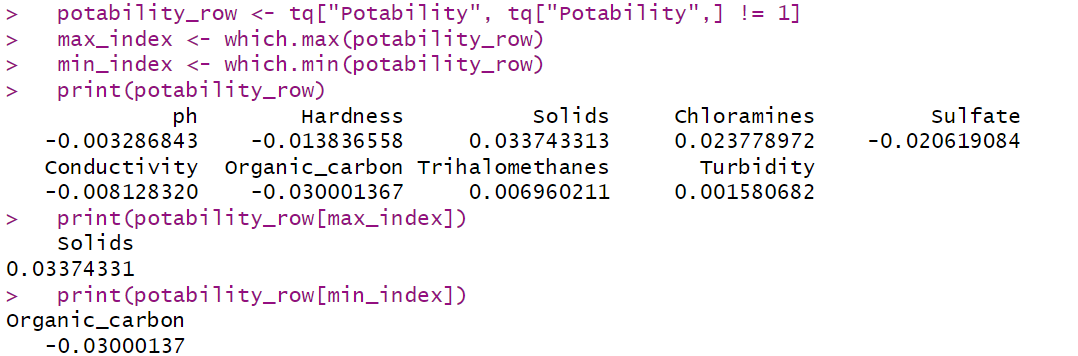






Hình 13: Đọc giá trị lớn nhất và giá trị nhỏ nhất của hệ số tương quan với biến đích là “Potability” 

Hình 14 : Kết quả giá trị lớn nhất và giá trị nhỏ nhất của hệ số tương quan với biến đích “Potability”:



**Nhận xét:**

Từ bảng số liệu, biểu đồ tương quan và sơ đồ tổng quát giữa từng cặp biến trong tập dữ liệu, ta có một vài nhận xét như sau:

Tất cả các hệ số trên đường chéo chính trong bảng dữ liệu bằng 1.

Ta nhận thấy ở hầu hết các sơ đồ sau khi loại bỏ đi đường chéo chính, các cặp biến đều có xu hướng tạo thành một đường nằm ngang. Điều này chứng tỏ các cặp biến có ít mối quan hệ tương quan với nhau. Khi ta quan sát các cặp biến với Potability với các biến khác thì bởi vì Potability chỉ có 2 giá trị 0 và 1 và đồ thị thể hiện rõ là 2 đường song song.

Từ việc xem xét các giá trị tương quan so với biến đích là biến Potability, ta nhận thấy các biến đều rất bé (trong đó biến có giá trị lớn nhất và nhỏ nhất lần lượt là 0.03374331 và -0.03000137).

**Kết luận:**

Qua quá trình phân tích các mối quan hệ hồi quy tuyến tính giữa các cặp biến thì ta có thể kết luận rằng, các cặp biến hầu như không có ảnh hưởng gì nhau và ngay cả với biến mục tiêu (biến Potability) cũng ít chịu ảnh hưởng bởi các cặp biến khác.

* + 1. **Hồi quy tuyến tính cho toàn bộ các biến độc lập:**

**Mục tiêu:**

Xét mô hình hồi quy tuyến tính bao gồm 10 biến trong đó biến Potability là một biến phụ thuộc vào tất cả các biến khác (đều là biến độc lập), bao gồm: ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic\_carbon, Trihalomethanes, Turbidity để kiểm tra sự phụ thuộc của biến Potability là một biến phụ thuộc với tất các biến khác.

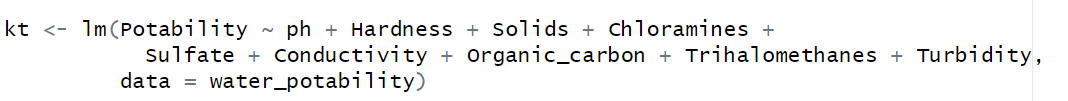
**Ý tưởng :**

Sử dụng lệnh lm() để tạo ra mô hình hồi quy tuyến tính với Potability là biến phụ thuộc, tất cả các biến còn lại đều là biến độc lập.

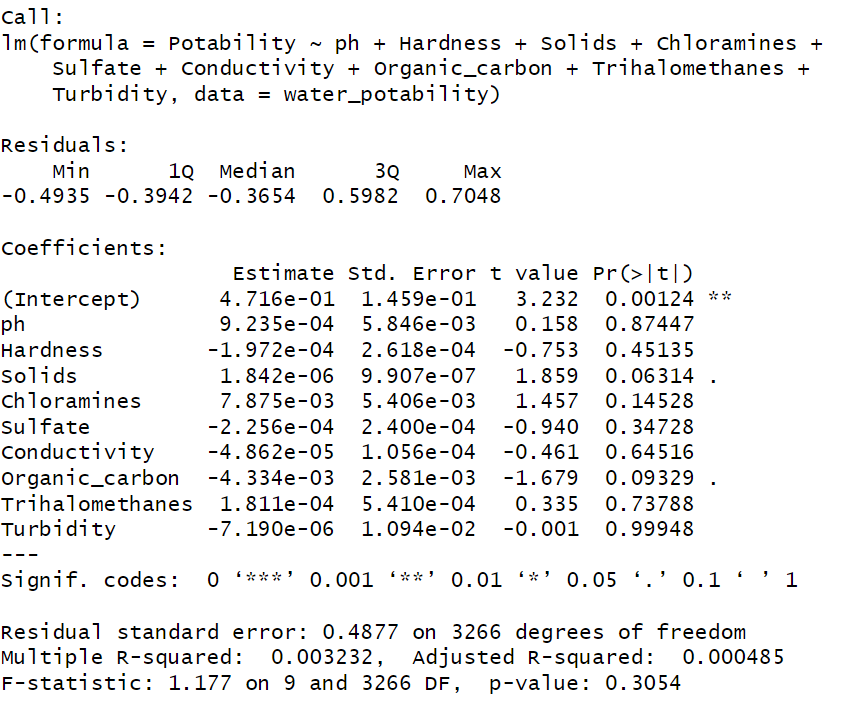
Sử dụng lệnh summary() để liệt kê các thông tin tính toán cần thiết để giải quyết vấn đề của bài toán.

**Hiện thực và kết quả:**

Hình 15 : Đoạn code và kết quả mô hình hồi quy tuyến tính cho toàn bộ biến độc lập.



Ta thu được kết quả sau:



**Nhận xét:**

Từ các giá trị thu được, ta có một vài nhận xét như sau:

Hệ số tương quan R (Adjusted R-squared) dùng để đo mức độ chặt chẽ của mô hình, càng về 1 càng chính xác và ngược lại. Kết quả thu được trong mô hình là 0.000485

Hệ số xác định R2  (Multiple R-squared) cho biết sự thay đổi là bao nhiêu phần trăm từ các giá trị biến độc lập. Kết quả thu được trong mô hình là 0.003232.

Trị số p (Pr(>|t|)) của các biến độc lập cho biết mức độ ảnh hưởng của từng biến, hệ số càng nhỏ thì mức độ ảnh hưởng càng mạnh. Ngoại trừ Intercept là biến chặn, biến có giá trị p lớn nhất là Turbidity với 0.99948 và biến có giá trị p nhỏ nhất là Solids với 0.06314.

**Kết luận:**

Qua các số liệu thu thập được, ta nhận thấy hệ số tương quan R rất bé và tiến sát về 0. Điều này cho thấy rằng mô hình có khả năng cao là không thể hiện được tính chặt chẽ và biến phụ thuộc Potability không phụ thuộc hoàn toàn vào các biến độc lập. Tuy nhiên để xét trong mô hình, biến nào là biến độc lập có khả năng ảnh hưởng cao nhất thì đó là biến Solids với 0.06314 cho thấy sự biến thiên của Solids có ảnh hưởng đáng kể đến việc nước đó có uống được hay không.

* + 1. **Hồi quy tuyến tính cho từng tổ hợp biến có trị số p mạnh:**

**Mục tiêu:**

Xét mô hình hồi quy tuyến tính bao gồm các tổ hợp biến trong đó biến Potability là một biến phụ thuộc và các tổ hợp biến bao gồm những biến có trị số p mạnh nhất để cải thiện chỉ số Hệ số tương quan R. Do đó, nhóm tiến hành xét trên tổ hợp từ 1 đến 5 biến có trị số p mạnh nhất.

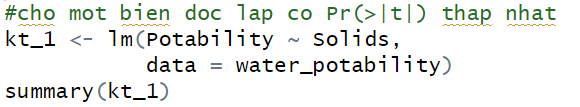
**Ý tưởng:**

Tương tự như mô hình hồi quy tuyến tính cho toàn bộ biến độc lập.

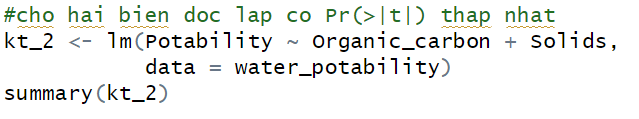
**Hiện thực và kết quả:**

Hình 16: Đoạn code và kết quả mô hình hồi quy tuyến tính cho từng tổ hợp biến

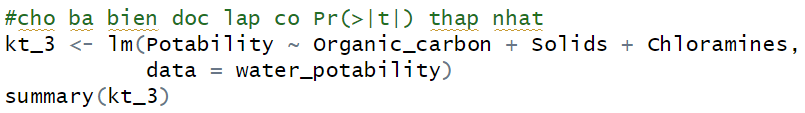
Một biến:



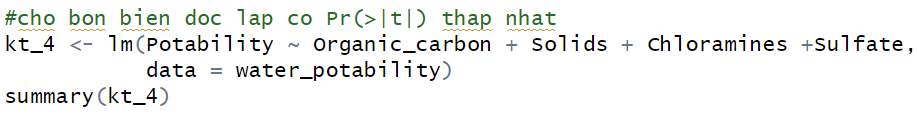
Hai biến:



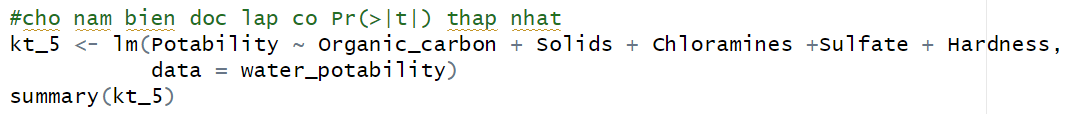
Ba biến:



Bốn biến:

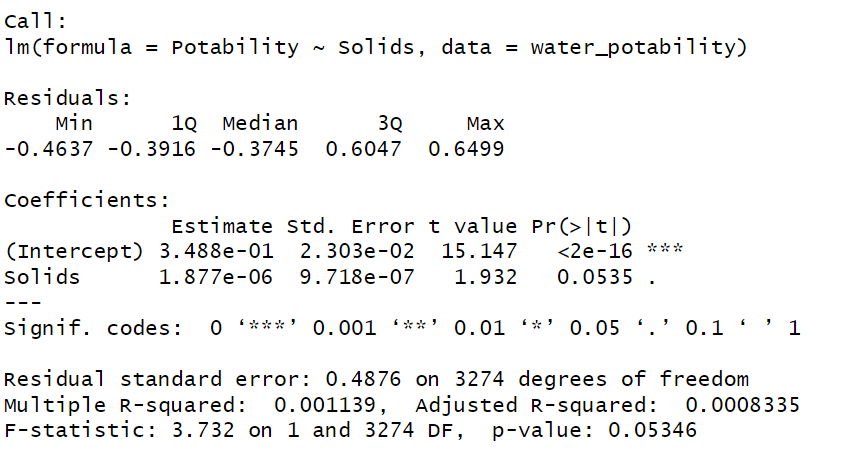


Năm biến:

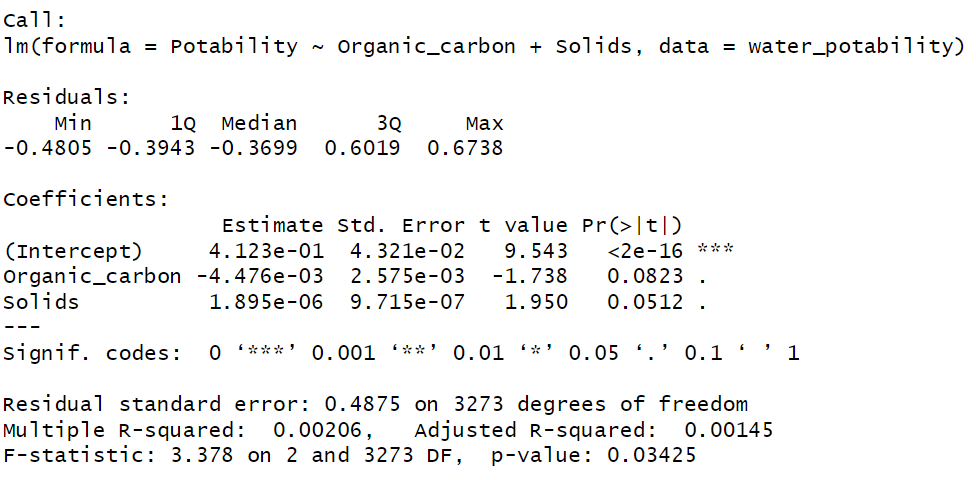


Ta thu được kết quả:

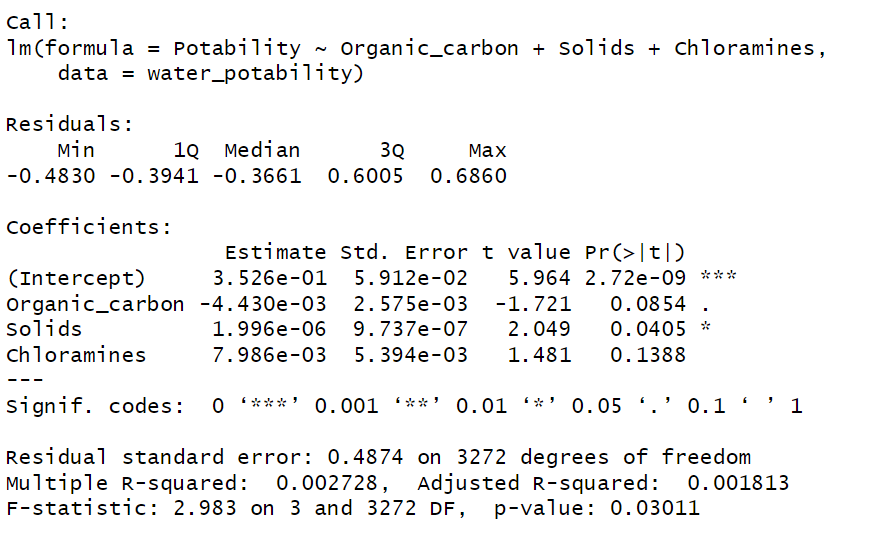
Một biến:



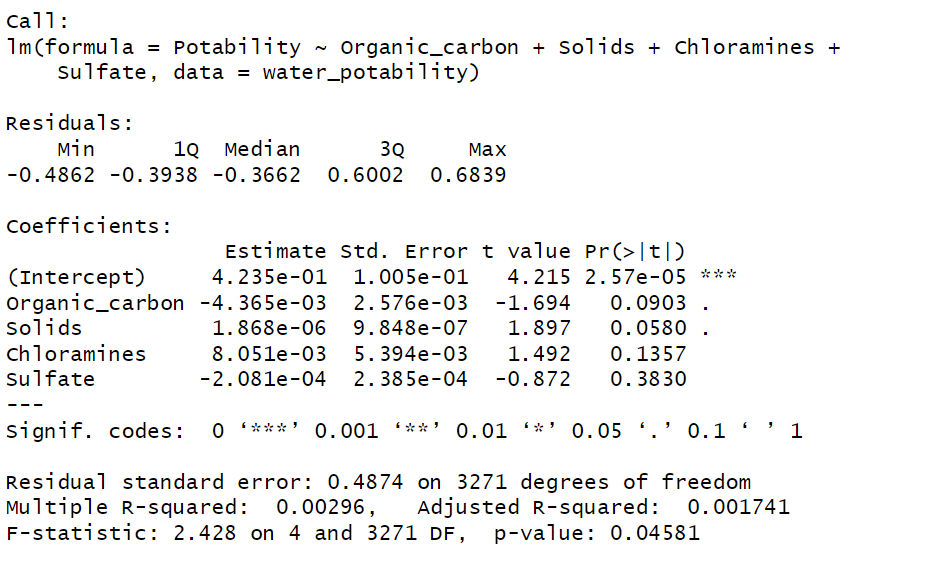
Hai biến:



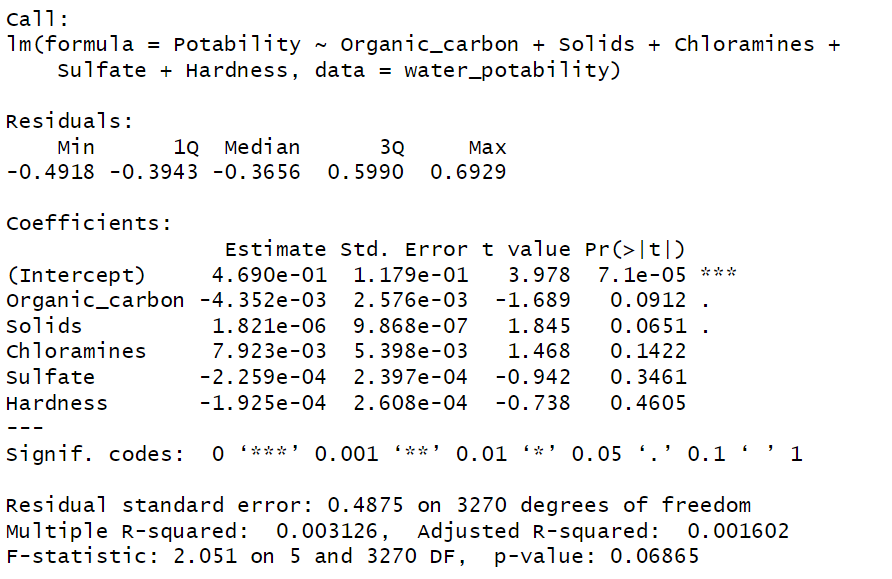
Ba biến:



Bốn biến:



Năm biến:



**Nhận xét:**

Từ các giá trị thu được, ta có một vài nhận xét như sau:

Hệ số tương quan R (Adjusted R-squared) của mô hình 1 biến đến 5 biến độc lập thu được kết quả lần lượt là 0.0008335, 0.00145, 0.001813, 0.001741 và 0.001602.

Hệ số xác định R2  (Multiple R-squared) của mô hình 1 biến đến 5 biến độc lập thu được kết quả lần lượt là 0.001139, 0.00206, 0.002728, 0.00296 và 0.003126.

Trị số p (Pr(>|t|)) của các biến độc lập ở tất cả các mô hình với mức độ ảnh hưởng mạnh yếu không đổi về mặt thứ tự so với mô hình cho toàn bộ biến độc lập nhưng các giá trị p đã bị thay đổi ít nhiều. Các tổ hợp biến có trị số p có xu hướng giảm từ tổ hợp 1 đến 3 biến và tăng từ tổ hợp 3 đến 5 biến. Và trong tất cả các mô hình này, Solids luôn luôn giữ trị số mạnh nhất.

**Kết luận:**

Từ những nhận xét trên, ta nhận thấy ở các tất cả các tổ hợp biến mà nhóm xét, chỉ số hệ số tương quan R đã được cải thiện đáng kể. Mô hình hồi quy tuyến tính cho 3 biến cho hệ số cao nhất với 0.00145. Tuy nhiên, từng đó vẫn là chưa đủ khi mà những giá trị này vẫn sát về phía 0 cho thấy dù có xét các tổ hợp biến, mô hình tuy đã chặt chẽ hơn một chút nhưng vẫn là không đáng kể và biến có ảnh hưởng nhất vẫn là biến Solids và ảnh hưởng mạnh nhất là trong mô hình hồi quy tuyến tính 3 biến Solids, Organic\_carbon và Chloramines.

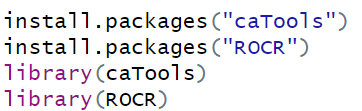
* 1. **Hồi quy Logistic**

**Mục tiêu:**

Tương tự như mô hình hồi quy tuyến tính, ta sẽ sử dụng biến Potability là một biến thụ thuộc và tất cả các biến còn lại là biến độc lập. Từ đây, ta sẽ xây dựng mô hình và đánh giá hiệu suất thông qua thống kê mô tả, dự đoán tỷ lệ chính xác trên tập huấn luyện và tập kiểm tra và dựa trên confusion matrix (ma trận nhầm lẫn) và AUC.

**Ý tưởng:**

Ta sẽ sử dụng các thư viện sau:



Sử dụng lệnh split để chia tập dữ liệu thành hai tập train và test với tỷ lệ 80:20 nhằm đảm bảo cho tập train đủ lớn để có thể tổng quát hóa dữ liệu tốt và tập test cũng cần đủ lớn để đảm bảo cho ra kết quả và đánh giá hiệu suất mô hình được chính xác hơn.

Sử dụng lệnh glm để xây dựng mô hình hồi quy logistic và đánh giá một vài tham số như p-value.

Sử dụng lệnh predict để dự đoán xác suất và table để lập confusion matrix.

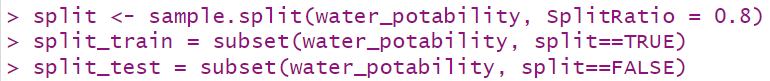
Sử dụng lệnh prediction, performance (với hai tham số là tpr và fpr) và plot để xây dựng đồ thị ROC.

Sử dụng lệnh performance với tham số là auc để đưa ra giá trị AUC và đánh giá mức độ dự báo của mô hình.

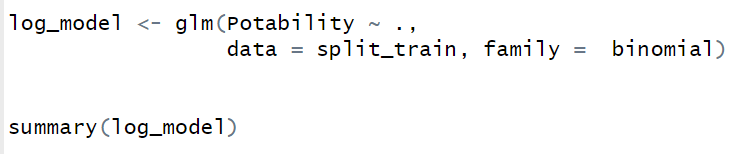
**Hiện thực và kết quả:**

Ta tiến hành hiện thực:

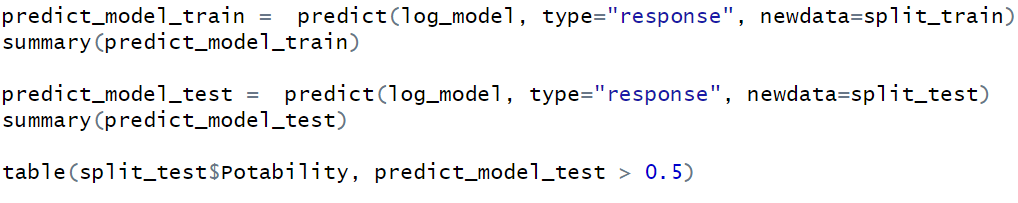
Hình 17: Chia dữ liệu thành train và test:



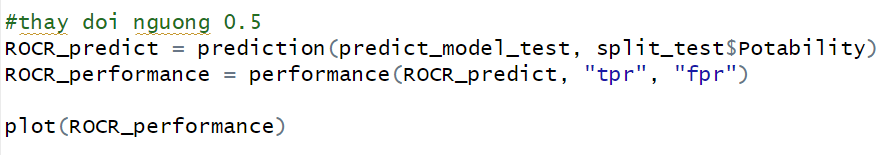
Hình 18: Xây dựng mô hình:



Hình 19: Đánh giá hiệu suất bằng confusion matrix:

****

Hình 20: Vẽ sơ đồ ROC:

****

Hình 21: Đoạn code và kết quả xem xét giá trị AUC:

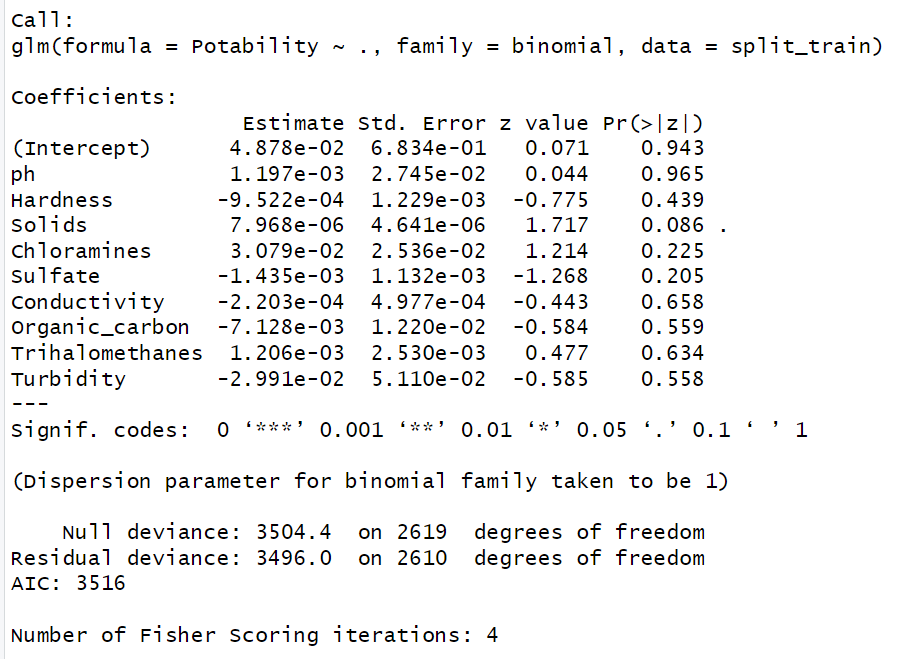
****

Ta thu được kết quả:

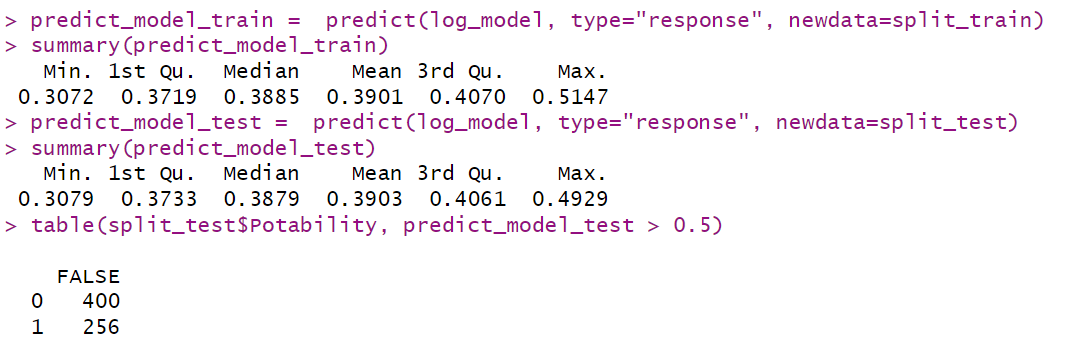
Chia dữ liệu thành train và test:



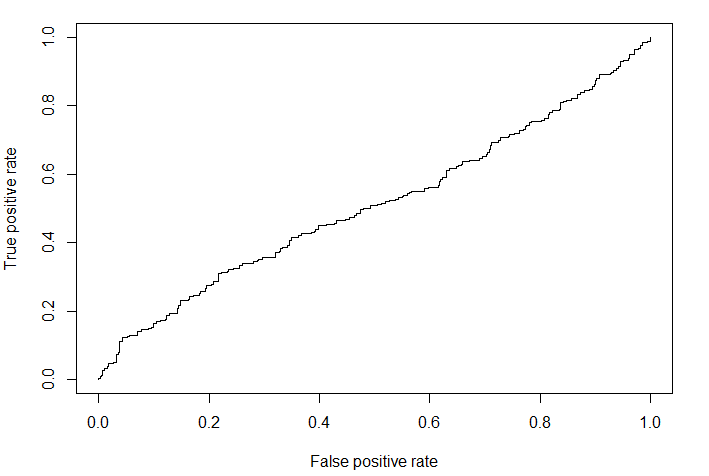
Hình 22: Xây dựng mô hình:



Hình 23: Đánh giá hiệu suất bằng confusion matrix:

****

Hình 24: Vẽ sơ đồ ROC:

****

Hình 25: Xem xét giá trị AUC:



**Nhận xét:**

Từ việc xây dựng mô hình hồi quy tuyến tính, ta nhận thấy rằng các giá trị p-value (Pr( > |z|)) khá cao và trên 0.05 với giá trị nhỏ nhất là 0.086 và giá trị lớn nhất là 0.965 tức là gần chạm đến 1.

Từ confusion matrix, ta nhận thấy rằng trong một tập mẫu test thì ta có tỷ lệ dự đoán là false – 0: 400 (tức là dự đoán không uống được và trên thực tế là không uống được) và false – 1: 256 (tức là dự đoán không uống được nhưng trên thực tế là uống được).

Từ đồ thị ROC, ta nhận thấy đây là một đường ROC có dạng hơi tuyến tính và bị lệch một góc 45 độ.

Giá trị AUC có phạm vi từ 0.5 đến 1, giá trị này càng lớn thì mô hình càng tốt và dự đoán được chất lượng trên thực tế. Và kết quả thu được là 0.5112402.

**Kết luận:**

Từ những số liệu nhận xét trên, ta nhận thấy rằng đối với mô hình hồi quy logistic, tất cả các biến độc lập hầu như không ảnh hưởng một cách đáng kể lên biến thuộc (Potability). Mặc dù trong quá trình xây dựng confusion test, ta thu được kết quả dự đoán tương đối khả quan (xấp xỉ 60.98% độ chính xác), tuy nhiên đồ thị ROC và giá trị AUC đã đưa ra một điều ngược lại khi mà đồ thị dạng tuyến tính và góc gần xấp xỉ 45 độ chứng tỏ rằng mô hình có xu hướng dự đoán ngẫu nhiên khi thay đổi giá trị ngưỡng. Sở dĩ ta thu được kết quả 60.98% là do ở gần gốc đồ thị, đồ thị có xu hướng hơi cong lên làm tăng khả năng đồ thị dự đoán đúng ở một vài ngưỡng nhất định. Ngoài ra giá trị AUC như đã nói ở trên, nếu giá trị đồ thị có xu hướng gần tới 0.5 tức là mô hình đang có xu hướng dự đoán ngẫu nhiên, thiếu chính xác và ngược lại khi ta tiến tới 1. Với giá trị 0.5112402, giá trị này không được tốt lắm và khó có thể dự đoán được chính xác trong thực tế.

1. **Thảo luận và mở rộng**

## **6.1. Hồi quy logistic**

Ưu điểm:

* Dùng trong trường hợp biến phụ thuộc là biến phân loại, vậy nên kết quả cũng là biến phân loại có thể giúp xác định cụ thể khả năng xảy ra. Mô hình này được sử dụng rộng rãi trong nghiêm cứu y học, kinh tế, xẫ hội và nhiều lĩnh vực để phân tích và dự đoán.
* Dễ thực hiện hơn nhiều so với các phương pháp khác, không chỉ cho phép đo lường mức độ liên quan của một biến độc lập mà còn cho chúng ta biết về hướng của mối quan hệ (tích cực hay tiêu cực).

Nhược điểm:

* Không dự đoán được kết quả liên tục, giá định tính tuyến giữa biến dự đoán (phụ thuộc) và biến dự đoán (độc lập).
* Yêu cầu số liệu thu thập nhiều dữ liệu, có thể tính không chính xác nếu kích thước cỡ mẫu quá nhỏ.

## **6.2. Mô hình Decision tree**

Dùng cho các mô hình đã xây dựng để dự đoán các giá trị mới

Hình 26: Đồ thị dự đoán của Decision tree

A diagram of a network

Description automatically generated

Nhận xét: Mỗi nút trong (internal node) tương ứng với một biến, đường nối giữa nó với nút con thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Từ đó kết luận chung rằng khả năng dự đoán đúng của mô hình đạt tỷ lệ cao

Hình 27: Dữ liệu dự đoán từ mô hình Decision Tree được đưa vào cột predicted\_2.

A screenshot of a computer

Description automatically generated

Hình 28: Kết quả dự đoán khả năng uống được hay không từ mô hình Decision Tree trên tập test.

A number and a number

Description automatically generated with medium confidence

Bảng so sánh dữ liệu quan sát so với dữ liệu dự đoán từ mô hình Decision Tree

Dòng Observation là dữ liệu quan sát từ thực tế, dòng Predicted\_value\_2 là ước tính từ mô hình.

Tạo bảng ma trận nhầm lẫn (confusion matrix) để biểu thể hiện số lượng các quan sát được phân loại đúng hoặc sai bởi mô hình Decision Tree, so với các thực tế của chúng.

A number on a white background

Description automatically generated

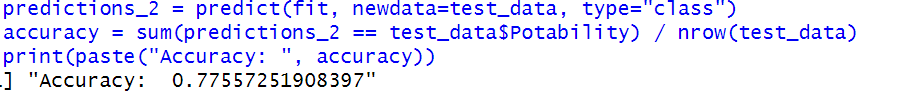
Ma trận nhầm lẫn trên code R.

Ta cụ thể hóa ma trận nhầm lẫn như sau:

|  |  |  |
| --- | --- | --- |
|  | Dự báo = 0 | Dự báo = 1 |
| Thực tế = 0 | 369 | 31 |
| Thực tế = 1 | 116 | 139 |

Bảng tần số của ma trận nhầm lẫn.

Tính toán độ chính xác của mô hình Decision Tree:



Nhận xét: Mô hình Decision Tree có độ chính xác là 78%, đây có thể xem là một hiệu suất khá lớn. Như vậy mô hình Decision Tree là một trong những mô hình có độ tin cậy cao.

**TÀI LIỆU THAM KHẢO**

Link đề bài: <https://www.kaggle.com/datasets/adityakadiwal/water-potability>

Link Code:

<https://drive.google.com/drive/folders/15ApfIAeyNGErvJy8xo7pz8S9imgPPFRx?usp=sharing>

Tài liệu tham khảo:

1. Nguyễn Đình Huy (2019), Giáo trính xác suất thống kê, Nhà xuất bản ĐHQG TPHCM
2. <https://www.kaggle.com/code/gcmadhan/water-quality-prediction-76-h2o-80-accuracy>
3. <https://www.kaggle.com/code/ihsncnkz/logistic-regression>
4. <https://www.kaggle.com/code/jcraggy/water-potability>
5. <https://www.kaggle.com/code/konnorherbst/water-quality-eda-quick-models>