ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KĨ THUẬT HÓA HỌC

BÁO CÁO BÀI TẬP LỚN

Môn học: XÁC SUẤT VÀ THỐNG KÊ

Đề tài: ỨNG DỤNG MÔ HÌNH HỒI QUY LOGISTIC ĐỂ DỰ ĐOÁN KHẢ NĂNG UỐNG NƯỚC

Lớp: L13 **Nhóm: 19**

Giảng viên hướng dẫn: Phan Thị Hường

Danh sách sinh viên:

Số thứ tự	Họ và tên	MSSV	Điểm
1	Nguyễn Thị Thanh Tuyền	2213830	0
2	Đoàn Mỹ Uyên	2213901	0
3	Phạm Thị Thanh Uyên	2213907	0
4	Hà Hoàng Vũ	2213990	0
5	Nguyễn Thị Việt Vương	2214016	0

Thành phố Hồ Chí Minh – Tháng 11/2023

Tỉ lệ đóng góp:

Số thứ tự	Họ và tên	MSSV	Tỉ lệ công việc	Tỉ lệ hoàn thành	Chữ ký
1	Nguyễn Thị Thanh Tuyền	2213830	20%	100%	
2	Đoàn Mỹ Uyên	2213901	20%	100%	
3	Phạm Thị Thanh Uyên	2213907	20%	100%	
4	Hà Hoàng Vũ	2213990	20%	100%	
5	Nguyễn Thị Việt Vương	2214016	20%	100%	

Nhóm trưởng: *Nguyễn Thị Thanh Tuyền*

Đánh giá của giảng viên:

Số thứ tự	Họ và tên	MSSV	Đánh giá
1	Nguyễn Thị Thanh Tuyền	2213830	
2	Đoàn Mỹ Uyên	2213901	
3	Phạm Thị Thanh Uyên	2213907	
4	Hà Hoàng Vũ	2213990	
5	Nguyễn Thị Việt Vương	2214016	

MỤC LỤC

Giới thiệu

Nước là một hợp chất hóa học được cấu tạo từ một nguyên tử Oxi và 2 nguyên tử Hydro. Mặc dù là một hợp chất vô cơ nhưng nước chiếm tới 75-80% trọng lượng cơ thể. Vì vậy việc cung cấp nước một cách đầy đủ cho cơ thể là một yêu cầu tất yếu. Đối với cơ thể sống nói chung, nước có các vai trò như: nước là dung môi hòa tan các chất; là môi trường diễn ra các phản ứng sinh hóa trong cơ thể; nước cung cấp nguồn chất khoáng cho cơ thể, giúp vận chuyển các chất cần thiết để nuôi dưỡng tế bào;… Chính vì thế, nhu cầu uống nước của cơ thể là một vấn đề được quan tâm và tập trung nghiên cứu trong xã hội hiện nay. Theo xu hướng ngày càng nâng cao chất lượng cuộc sống, vấn đề uống nước không chỉ quan tâm đến số lượng mà bên cạnh đó là yếu tố chất lượng nguồn nước. Về lý thuyết để đánh giá chất lượng một nguồn nước có thể uống được hay không phải dựa vào rất nhiều quy trình phức tạp nhờ và hệ thống máy móc xử lí và phân tích trực tiếp nguồn nước. Tuy nhiên về cơ bản người ta có thể đánh giá một cách tương đối khả năng uống được của một nguồn nước thông qua các giá trị chỉ số đặc trưng của nó. Ví dụ một vài chỉ số quen thuộc như độ pH, độ cứng, độ dẫn điện,... và từ các giá trị đó, thông qua xây dựng các thuật toán, tính toán để đưa ra kết luận nguồn nước có khả năng uống được hay không. Căn cứ vào những yếu tố đó, đồng thời phục vụ cho yêu cầu nghiên cứu, báo cáo bài tập lớn môn học XÁC SUẤT THỐNG KÊ, nhóm chúng em đã thống nhất lựa chọn đề tài nghiên cứu: Ứng dụng mô hình Hồi quy Logistic để dự đoán khả năng uống được của nước.

1.Tổng quan dữ liệu

1.1. Ngữ cảnh của dữ liệu

Tiếp cận nguồn nước uống an toàn là điều cần thiết cho sức khỏe, một quyền cơ bản của con người và là một phần của 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 cấp nước và vệ sinh có thể mang lại lợi ích kinh tế ròng vì việc giảm các tác động tiêu cực đến sức khỏe và chi phí chăm sóc sức khỏe lớn hơn chi phí thực hiện các biện pháp can thiệp. Dữ liệu Water Quality giúp đánh giá được chất lượng nước của các vùng khác nhau của các nước khác nhau.

1.2. Tổng quan về dữ liệu

Tệp dữ liệu Water Quality mô tả các đặc trưng được sử dụng thông dụng để đánh giá mức độ an toàn đối với con người của nước. Bao gồm: pH value, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity, và Potability. Bao gồm chất lượng nước của 3276 vùng khác nhau của các nước khác nhau. Trong đó:

pH value:

pH là một thông số quan trọng trong việc đánh giá cân bằng axit-bazơ của nước. Nó cũng là chỉ số dùng để đánh giá về tình trạng axit hoặc kiềm của trạng thái nước. WHO đã khuyến nghị giới hạn pH tối đa cho phép để nước có thể sử dụng được là từ 6,5 đến 8,5. Trong các điều tra hiện nay theo tiêu chuẩn của WHO là từ 6,52 đến 6,83.

Hardness (độ cứng):

Độ cứng chủ yếu là do các ion Ca^{2+} và Mg^{2+} trong nước gây ra. Những ion này được hòa tan từ các trầm tích địa chất mà nước di chuyển qua nó. Khoảng thời gian nước tiếp xúc với vật liệu tạo độ cứng, giúp xác định độ cứng trong nước thô. Độ cứng ban đầu được định nghĩa là khả năng của nước làm kết tủa xà phòng do Ca^{2+} và Mg^{2+} gây ra. Nước càng có nhiều Ca^{2+} và Mg^{2+} thì độ cứng càng cao. Độ cứng của nước có thể gây ra các vấn đề trong việc sử dụng nước trong các hệ thống cung cấp nước, thiết bị gia dụng và các quy trình công nghiệp.

Solids (Total dissolved solids – TDS):

Nước có khả năng hòa tan nhiều loại khoáng chất vô cơ và hữu cơ hoặc các muối như kali, canxi, natri, bicarbonat, clorua, magie, sunfat, v.v. Những khoáng chất này tạo ra mùi vị không mong muốn và màu sắc loãng khi xuất hiện của nước. TDS là một chỉ số đo lường lượng những chất rắn hòa tan này trong nước. Nước có giá trị TDS cao chứng tỏ nước có độ khoáng hóa cao, nhưng đồng thời cũng có thể chứa các chất có thể gây hại nếu nồng độ quá mức. Mức TDS cũng có thể ảnh hưởng đến vị giác và chất lượng của nước. Nước có TDS thấp có thể có hương vị nhạt nhòa, trong khi nước có TDS cao thường có hương vị mạnh mẽ hơn. Giới hạn tối thiểu đối với TDS là 500 mg/l và giới hạn tối đa là 1000 mg/l.

Chloramines:

Clo và Chloramine (một loại hợp chất hóa học chứa cả Clo và Ammonia) là chất khử trùng chính được sử dụng trong hệ thống nước công cộng. Chloramine thường được hình thành khi Ammonia được thêm vào clo để xử lý nước uống, chúng được tạo ra khi Clo, một chất khử trùng phổ biến trong quá trình xử lý nước, phản ứng với Ammonia hoặc các chất hữu cơ nitrogen khác trong nước. Tuy nhiên, chloramines cũng có thể gây ra một số vấn đề, ví dụ như mùi kháng khuẩn, tác động đối với một số quá trình sản xuất và nếu nước chứa chloramines được sử dụng trong hồ bơi có thể tạo ra các hợp chất phát tán khó chịu. Do đó, việc kiểm soát và giảm lượng chloramines trong nước cung cấp là quan trọng để đảm bảo chất lượng nước sạch và an toàn. Mức clo lên tới 4 miligam mỗi lít (mg/L hoặc 4 phần triệu (ppm)) được coi là an toàn trong nước uống.

Sulfate:

Sulfate là những chất tự nhiên có mặt trong khoáng sản, đất, và đá. Chúng xuất hiện trong không khí xung quanh, nước ngầm, cây cỏ, và thực phẩm. Mặc dù sulfate không gây nguy hiểm đối với sức khỏe ở mức độ thấp, nhưng ở mức độ cao có thể tạo ra một số tác động không mong muốn, như tạo ra một hương vị đặc trưng trong nước và ảnh hưởng đến chất lượng nước uống. Việc sử dụng chính của sulfate là trong công nghiệp hóa chất. Nồng độ sulfate trong nước biển là khoảng 2.700 miligam trên mỗi lít (mg/L). Trong hầu hết các nguồn nước ngọt, nồng độ này dao động từ 3 đến 30 mg/L, mặc dù có một số khu vực địa lý có nồng độ cao hơn (1000 mg/L).

Conductivity (Độ dẫn điện):

Nước tự nhiên thường chứa các chất hòa tan, chủ yếu là muối khoáng và các ion khác. Khi nồng độ các ion tăng lên, khả năng dẫn điện của nước cũng tăng theo. Do đó, đo lường độ dẫn điện có thể được sử dụng để ước lượng lượng chất rắn hòa tan trong nước. Nhìn chung, lượng chất rắn hòa tan trong nước xác định khả năng dẫn điện của nó. Dẫn điện (EC) thực sự đo lường quá trình ion hóa của một dung dịch, cho phép nó truyền dòng điện. Theo tiêu chuẩn của Tổ chức Y tế Thế giới (WHO), giá trị EC không nên vượt quá 400 µS/cm.

Organic Carbon:

Organic carbon trong nước là một thành phần quan trọng của chất hữu cơ có nguồn gốc từ sinh vật hoặc thực vật. Organic carbon bao gồm carbon hữu cơ không hòa tan và hòa tan trong nước. Carbon hữu cơ tổng cộng (TOC) trong nguồn nước đến từ quá trình phân hủy của chất hữu cơ tự nhiên (NOM) cũng như từ nguồn gốc tổng hợp. TOC đo lường tổng lượng carbon trong các hợp chất hữu cơ trong nước tinh khiết. Theo Cơ quan Bảo vệ Môi trường Hoa Kỳ (US EPA), giá trị TOC trong nước đã được xử lý/để uống nên dưới 2 mg/L và trong nguồn nước sử dụng cho quá trình xử lý nước không nên vượt quá 4 mg/L.

Trihalomethanes:

Trihalomethanes (THMs) là những hợp chất hóa học có thể xuất hiện trong nước đã được xử lý bằng clo. Nồng độ của THMs trong nước uống thường biến đổi tùy thuộc vào

mức độ vật liệu hữu cơ trong nước, lượng clo cần thiết để xử lý nước và nhiệt độ của nước đang được xử lý. Mức độ THMs lên đến 80 ppm được coi là an toàn trong nước uống.

Turbidity (độ đục):

Turbidity trong nước đề cập đến mức độ đục của nước, tức là mức độ mất khả năng nhìn thấy của nước do sự hiện diện của các hạt rắn, bùn, tảo, vi khuẩn, và các chất lẫn vào nước. Độ đục là một đặc điểm quan trọng để đánh giá chất lượng nước và có thể là một chỉ số của sự ô nhiễm. Đây là một đo lường về khả năng phát quang của nước và thử nghiệm này được sử dụng để chỉ ra chất lượng của nước thải liên quan đến chất hữu cơ tan trong nước. Giá trị độ đục trung bình theo khuyến nghị của WHO là 5,00 NTU.

Potability:

Khả năng uống được của nước cho biết được nước đó có an toàn cho con người hay khôn. Dựa vào các đặc trưng trên, người ta có thể đánh giá được khả năng uống được của nước thông qua việc đo lường các đặc trưng có đó nằm trong mức cho phép của nước có thể uống được theo khuyến nghị của WHO. Trong tệp dữ liệu này "1" có nghĩa là nước đó có thể uống được và "0" có nghĩa là không uống được.

2.Kiến thức nền

2.1. Hồi quy tuyến tính bội

2.1.1. Mô hình hồi quy tuyến tính bội

Hồi quy tuyến tính bội là mô hình được mở rộng từ mô hình hồi quy tuyến tính đơn, trong đó không phải chỉ một mà nhiều biến giải thích có thể được sử dụng để dự đoán giá trị của biến phụ thuộc.

Mô hình hồi quy tuyến tính bội có dạng tổng quát như sau:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

Trong đó:

Y: biến phụ thuộc

x_i:biến giải thích, $i = 1, \dots, k$

β_0: hệ số chặn

β_i: hệ số góc, $i = 1, \dots, k$

ϵ : thành phần sai số trong mô hình và tuân theo phân phối chuẩn

Các hệ số hồi quy β_0, β_i $(i = 1, \dots, k)$, sẽ miêu tả cho sự thay đổi kỳ vọng của biến phụ thuộc Y với mỗi sự thay đổi của biến x_i $(i = 1, \dots, k)$ khi các biến hồi quy còn lại x_j $(i \neq j)$ được giữ cố định.

2.1.2. Phương trình hồi quy tổng thể

Với Y là biến phụ thuộc vào $x_1, x_2, \dots, x_k$ là các biến độc lập, Y là ngẫu nhiên có một phân phối xác suất nào đó. Tồn tại $E(Y|x_1, x_2, \dots, x_k) = f_Y(x_1, x_2, \dots, x_k)$ là hàm hồi quy tổng thể của biến phụ thuộc Y theo biến giải thích $x_1, x_2, \dots, x_k$.

Nghĩa là hàm hồi quy của Y theo $x_1, x_2, \dots, x_k$ chính là kỳ vọng có điều kiện của Y đối với x_j, có dạng:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

Trong đó:

β_0: hệ số tung độ gốc (hệ số chặn)

β_1: hệ số góc của biến Y theo biến x_1 giữ các biến $x_1, x_2, \dots, x_k$ không đổi

β_2: hệ số góc của biến Y theo biến x_2 giữ các biến $x_1, x_3, \dots, x_k$ không đổi

β_k: hệ số góc của biến Y theo biến x_k giữ các biến $x_1, x_2, \dots, x_{k-1}$ không đổi

ϵ_i: thành phần ngẫu nhiên, có phân phối chuẩn với $E(\epsilon_i) = 0, Var(\epsilon_i) = \sigma^2$

2.1.3. Phương trình hồi quy mẫu

Do không thể biết được tổng thể, nên chúng ta không biết được giá trị trung bình tổng thể của biến phụ thuộc là đúng ở mức độ nào. Vì thế chúng ta phải dựa vào dữ liệu mẫu để ước lượng.

Ta có hàm hồi quy mẫu tổng quát được viết dưới dạng như sau:

$$\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$$

Trong đó:

$\widehat{Y_i}$: là giá trị ước lượng của biến phụ thuộc Y_i căn cứ vào mô hình hồi quy

$\hat{\beta}_j$: là giá trị ước lượng cho hệ số hồi quy β_j căn cứ vào dữ liệu mẫu,

$(j = 0,1, \dots k)$

2.1.4. Sử dụng phương pháp bình phương cực tiểu để xác định hệ số hồi quy

Phương trình hồi quy tuyến tính có dạng:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

Để xác định được các hệ số hồi quy β_j $(j = 0,1,2, \dots, k)$ ta sử dụng hàm phương pháp bình phương cực tiểu, và hàm bình phương cực tiểu được xác định:

$$L = \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{n} \beta_j . x_{ij})^2$$

Tìm giá trị nhỏ nhất của hàm L ta sẽ tìm được các hệ số hồi quy mẫu cho phương trình hồi quy.

Trong khi xây dựng mô hình hồi quy tuyến tính bội, ta cần kiểm tra các giả thuyết như: hàm hồi quy là hàm tuyến tính theo các tham số, sai số ngẫu nhiên độc lập với nhau tuân theo phân phối chuẩn với kỳ vọng bằng 0 và phương sai là σ^2.

2.1.5. Kiểm định mức độ phù hợp của mô hình

a) Hệ số xác định bội

Để xác định được phần biến thiên trong biến phụ thuộc được giải thích bởi mối liên hệ giữa biến phụ thuộc và tất cả các biến độc lập trong mô hình, người ta đi xác định hệ số xác định bội R^2 ($0 \leq R^2 \leq 1$). Hệ số xác định bội sẽ giải thích trong 100% sự biến động của Y so với trung bình của nó thì có bao nhiêu % là do biến các biến x_j, gây ra, công thức để tính hệ số xác định bội là:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Trong đó:

SST (Sum of Squares Total): là tổng bình phương tất cả các sai lệch giữa các giá trị của biến phụ thuộc và giá trị trung bình

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

SSR (Sum of Squares in Regression): là tổng bình phương của tất cả các sai lệch giữa các giá trị của biến phụ thuộc Y nhận được từ hàm hồi quy mẫu và giá trị trung bình của chúng.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SSE (Sum of Squares for Error): dùng để đo sự chênh lệch giữa từng giá trị quan sát với giá trị dự đoán. SSE được xem như sai số do những yếu tố ngoài X hoặc do lấy mẫu ngẫu nhiên.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SST - SSR$$

R^2 cao nghĩa là sự biến động của biến phụ thuộc Y so với trung bình của nó do biến X gây ra càng cao

Nếu $R^2 = 1$ nghĩa là đường hồi quy giải thích cho 100% sự thay đổi của Y

Nếu $R^2 = 0$ nghĩa là mô hình không đưa ra thông tin nào về sự thay đổi của biến phụ thuộc Y.

b) Hệ số xác định điều chỉnh

Do việc đưa thêm biến độc lập vào mô hình sẽ luôn làm gia tăng hệ số xác định R^2, thậm chí ngay cả khi biến độc lập được đưa vào không có mối liên hệ hoặc có mối liên hệ không đáng kể với biến phụ thuộc. Vì thế khi số biến độc lập tăng lên chắc chắn R^2 sẽ luôn tăng lên, tuy nhiên mỗi biến độc lập được thêm vào sẽ làm mất đi một bậc tự do. Sự gia tăng trong R^2 có thể không bù đắp được thiệt hại do mất thêm bậc tự do khi thêm biến, thế nhưng ${R^2}_{adj}$ lại có xét đến và điều chỉnh giá trị của ${R^2}_{adj}$ theo nó một cách phù hợp. ${R^2}_{adj}$ sẽ luôn bé hơn R^2. Khi một biến độc lập được thêm vào không có đóng góp xứng đáng vào khả năng giải thích cho biến phụ thuộc thì ${R^2}_{adj}$ sẽ luôn giảm mặc dù R^2 thì tăng.

Hệ số xác định điều chỉnh là một đại lượng đo lường quan trọng khi số biến độc lập lớn một cách tương đối so với cỡ mẫu, nó tính đến mối liên hệ giữa cỡ mẫu và số biến, nếu số biến độc lập là khá lớn so với cỡ mẫu thì R^2 sẽ thổi phồng khả năng giải thích cho biến phụ thuộc của mô hình một cách giả tạo, thế nên người ta cần xài đến hệ số xác định điều chỉnh để đánh giá được chính xác hơn sự thay đổi của biến phụ thuộc Y vào biến độc lập X. Hệ số xác định điều chỉnh được xác định bằng công thức:

$$ {R^2}_{adj} = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SST}{n-1}} = 1 - (1-R^2)\left[\frac{n-1}{n-k-1}\right] $$

c) Đánh giá ý nghĩa toàn diện của mô hình

Kiểm định mức ý nghĩa của mô hình hồi quy là kiểm định để xác định xem có tồn tại mỗi quan hệ tuyến tính hay không giữa biến phụ thuộc Y_i và tập con của các biến độc lập $x_1, x_2, \dots, x_k$. Vì mô hình hồi quy mà chúng ta xây dựng là dựa trên dữ liệu của một mẫu lấy từ tổng thể vì vậy nó có thể bị ảnh hưởng của sai số lấy mẫu, chính vì thế mà chúng ta cần phải kiểm định ý nghĩa thống kê của mô hình. Như đã biết, hệ số xác định bội R^2 là đại lượng cho biết có bao nhiêu phần trăm biến thiên trong biến phụ thuộc có thể được giải thích bởi mô hình hồi quy, là một số thống kê trên mẫu có thể sử dụng để suy diễn về việc mô hình toàn diện có ý nghĩa về mặt thống kê hay không trong việc giải thích cho biến thiên của biến phụ thuộc. Với ý tưởng này, chúng ta đặt ra các giả thuyết kiểm định như sau:

$$ H_0: R^2 = 0 $$

$$ H_1: R^2 \neq 0 $$

Bản chất của giả thuyết H_0 này là mô hình hồi quy tuyến tính bội tổng thể mà chúng ta xây dựng thì các biến độc lập không giải thích được chút nào cho những biến thiên trong biến phụ thuộc. Tương tự, ta có thể đặt lại giả thuyết:

$$ H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0 $$

$$ H_1: Có\ ít\ nhất\ một\ hệ\ số\ \beta_j \neq 0 $$

Nếu ta bác bỏ giả thuyết H_0 đồng nghĩa với việc có ít nhất một trong những biến hồi quy $x_1, x_2, \dots, x_k$ có đóng góp đáng kể cho mô hình

Kiểm định thống kê cho giả thuyết H_0 được xác định bằng công thức:

$$F_0 = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$$

Chúng ta sẽ bác bỏ giả thuyết H_0 nếu như giá trị của kiểm định thống kê f_0 là lớn hơn $f_{\alpha,k,n-k-1}$.

2.1.6. Ước lượng khoảng tin cậy cho hệ số hồi quy

Trong mô hình hồi quy tuyến tính bội, người ta thường xác định khoảng tin cậy của các hệ số hồi quy để xác định được phân bố xác suất của các hệ số hồi quy β_j (j=0,1,..k). Việc xây dựng khoảng tin cậy cho các hệ số hồi quy này đòi hỏi sai số ϵ_i phải có phân phối chuẩn với kỳ vọng bằng 0 và phương sai là σ^2

$$\epsilon_i \sim N(0, \sigma^2)$$

Do đó, từ giả định của sai số ngẫu nhiên, suy ra biến phụ thuộc Y_i sẽ có phân phối chuẩn với kỳ vọng là $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$ và phương sai là σ^2. Vì ước lượng điểm cho các hệ số hồi quy $\hat{\beta}_j$ là một tổ hợp tuyến tính nên $\hat{\beta}_j$ sẽ có phân phối chuẩn với kỳ vọng là β_j phương sai là ma trận hiệp phương sai. Sau đó, mỗi thống kê:

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^{\,2} C_{jj}}} \qquad j = 0,1,\dots,k$$

sẽ có phân phối T với bậc tự do là $n-k-1$

Trong đó:

$\hat{\sigma}^{\,2}$: là ước lượng điểm cho phương sai lỗi, và được xác định bằng

$$\hat{\sigma}^{\,2} = \frac{SSE}{n-k-1}$$

$\sqrt{\hat{\sigma}^{\,2} C_{jj}}$: là sai số chuẩn của hệ số hồi quy $\hat{\beta}_j$, kí hiệu $se(\hat{\beta}_j)$

Điều này dẫn đến với độ tin cậy $1-\alpha$ thì khoảng ước lượng cho hệ số hồi quy β_j, j=0,1,..,k cho mô hình hồi quy tuyến tính bội được đưa ra bởi:

$$\hat{\beta}_j - t_{\frac{\alpha}{2},n-k-1} se(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{\frac{\alpha}{2},n-k-1} se(\hat{\beta}_j)$$

2.1.7. Kiểm định hệ số hồi quy

Thông thường, khi xây dựng mô hình hồi quy tuyến tính bội, người ta thường quan tâm đến việc kiểm tra các giả thuyết về các hệ số hồi quy riêng lẻ. Các thử nghiệm như vậy sẽ hữu ích trong việc xác định giá trị tiềm năng của từng biến hồi quy. Ví dụ, mô hình có thể có hiệu quả hơn nếu bao gồm các biến bổ sung hoặc có thể là việc xóa bỏ một biến hồi quy

trong mô hình. Cũng giống như mô hình hồi quy tuyến tính đơn, kiểm định thống kê cũng yêu cầu sai số ngẫu nhiên ϵ_i có phân phối chuẩn với kỳ vọng bằng 0 và phương sai là σ^2.Giả thuyết để kiểm tra hệ số hồi quy riêng lẻ, giả định rằng:

$$H_0: \beta_j = \beta_{j0}$$

$$H_1: \beta_j \neq \beta_{j0}$$

Và thống kê kiểm định cho giả thuyết này là:

$$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j - \beta_{j0}}{se(\hat{\beta}_j)}$$

Giả thuyết H_0 sẽ bị bác bỏ nếu như $|t_0| > t_{\frac{\alpha}{2}, n-k-1}$.

Một trường hợp đặc biệt quan trọng của giả thuyết H_0, là thông thường ta sẽ kiểm định giả thuyết H_0, H_1 với $\beta_{j0} = 0$

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Nếu như giả thuyết H_0 không bị bác bỏ, điều này đồng nghĩa biến hồi quy x_j có thể được xóa bỏ khỏi mô hình. Việc thêm biến vào mô hình hồi quy luôn làm cho tổng bình phương của tất cả các sai lệch giữa các giá trị của biến phụ thuộc Y nhận được từ hàm hồi quy mẫu và giá trị trung bình của chúng (SSR) tăng lên và tổng bình phương của lỗi (SSE) giảm xuống (điều này giải thích tại sao hệ số xác định bội R^2 luôn tăng khi ta thêm biến hồi quy vào mô hình). Chúng ta phải xem xét rằng mức tăng trong SSR có đủ lớn khi ta thêm một biến mới vào hay không, hơn nữa việc thêm một biến mới không quan trọng vào mô hình có thể làm cho sai số ngẫu nhiên tăng lên. Điều này cho thấy rằng việc thêm một biến như vậy thực sự đã làm cho mô hình kém phù hợp hơn so với dữ liệu (Điều này lí giải tại sao hệ số xác định điều chỉnh ${R^2}_{adj}$ là biện pháp tốt hơn trong việc kiểm tra độ phù hợp của mô hình).Ngoài ra, còn có các giả thuyết kiểm định về phía trái hay phía phải cho các hệ số hồi quy riêng lẻ.

Đối với phía phải ta có các giả thuyết

$$H_0: \beta_j \leq \beta_{j0}$$

$$H_1: \beta_j > \beta_{j0}$$

Thống kê kiểm định cũng giống như kiểm định về hai phía. Và giả thuyết H_0 sẽ bị bác bỏ nếu như $t_0 > t_{\frac{\alpha}{2}, n-k-1}$. Việc bác bỏ giả thuyết H_0 đồng nghĩa với việc ta chấp nhận giả thuyết H_1, điều đó có nghĩa là x_j có tác động thuận đối với mô hình.

Tương tự cho kiểm định về bên trái, ta cũng đặt các giả thuyết:

$$H_0: \beta_j \geq \beta_{j0}$$

$$H_1: \beta_j < \beta_{j0}$$

Giả thuyết H_0 sẽ bị bác bỏ nếu như $t_0 < -t_{\frac{\alpha}{2},n-k-1}$. Điều này đồng nghĩa với x_j sẽ có tác động ngược đối với mô hình.

Giả thuyết	Giả thuyết H_0	Giả thuyết H_1	Miền bác bỏ
Hai phía	$\beta_j = \beta_{j0}$	$\beta_j \neq \beta_{j0}$	$\|t_0\| > t_{\frac{\alpha}{2},n-k-1}$
Phía phải	$\beta_j \leq \beta_{j0}$	$\beta_j > \beta_{j0}$	$t_0 > t_{\frac{\alpha}{2},n-k-1}$
Phía trái	$\beta_j \geq \beta_{j0}$	$\beta_j < \beta_{j0}$	$t_0 < -t_{\frac{\alpha}{2},n-k-1}$

2.2. Hồi quy logistic

2.2.1. Khái niệm

Hồi quy logistic là một mô hình thống kê được sử dụng chủ yếu để dự đoán xác suất của một biến phụ thuộc nhị phân. Nó thường được sử dụng trong các vấn đề phân loại, nơi mục tiêu là phân loại một quan sát vào một trong hai nhóm. Giống như tất cả các phân tích hồi quy, hồi quy logistic là một 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 biến nhị phân phụ thuộc và một hoặc nhiều biến độc lập.

2.2.2. Các loại hồi quy logistic

a. Hồi quy logistic nhị phân:

Hồi quy logistic nhị phân được sử dụng để dự đoán xác suất của kết quả nhị phân, chẳng hạn như có hoặc không, đúng hay sai hoặc 0 hoặc 1. Ví dụ: nó có thể được sử dụng để dự đoán liệu một khách hàng có rời bỏ hay không, liệu bệnh nhân có có bệnh hay không, hoặc khoản vay có được hoàn trả hay không.

b. Hồi quy logistic đa thức:

Hồi quy logistic đa thức được sử dụng để dự đoán xác suất của một trong ba kết quả trở lên có thể xảy ra, chẳng hạn như loại sản phẩm mà khách hàng sẽ mua, xếp hạng mà khách hàng sẽ đưa ra cho sản phẩm hoặc đảng phái chính trị mà một người sẽ bỏ phiếu.

c. Hồi quy logistic thông thường:

Được sử dụng để dự đoán xác suất xảy ra một kết quả theo thứ tự định trước, chẳng hạn như mức độ hài lòng của khách hàng, mức độ nghiêm trọng của bệnh hoặc giai đoạn ung thư.

2.2.3. Tại sao nên sử dụng hàm hồi quy logistic thay vì hồi quy tuyến tính

Theo định nghĩa của hàm hồi quy logistic, nó chỉ được sử dụng khi biến phụ thuộc của chúng ta là nhị phân và trong hồi quy tuyến tính, biến phụ thuộc này là liên tục.

Nếu chúng ta thêm một ngoại lệ vào tập dữ liệu của mình thì đường phù hợp nhất trong hồi quy tuyến tính sẽ thay đổi để phù hợp với điểm đó. Nghĩa là đường hồi quy ban đầu của biến phụ thuộc, khi được thêm vào một biến giải thích mới thì đường hồi quy đó sẽ thay đổi sao cho phù hợp với biến giải thích mới được thêm vào. Làm thay đổi kết quả đầu ra của biến phụ thuộc.

Một vấn đề khác với hồi quy tuyến tính là các giá trị dự đoán có thể nằm ngoài phạm vi. Chúng ta biết rằng xác suất có thể nằm trong khoảng từ 0 đến 1, nhưng nếu chúng ta sử dụng hồi quy tuyến tính thì xác suất này có thể vượt quá 1 hoặc xuống dưới 0.

Để khắc phục những vấn đề này, sử dụng Hồi quy logistic, chuyển đổi đường thẳng phù hợp nhất trong hồi quy tuyến tính thành đường cong chữ S bằng cách sử dụng hàm sigmoid, hàm này sẽ luôn cho các giá trị từ 0 đến 1.

2.2.4. Mô hình hồi quy logistic

a. Phương trình hồi quy logistic

Mô hình hồi quy logistic có dạng tổng quát như sau:

$$p(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}$$

Trong đó:

Y=1: xác suất xảy ra sự kiện của biến phụ thuộc

β_0: hệ số tung độ gốc (hệ số chặn)

β_1: hệ số góc của biến Y theo biến x_1 giữ các biến $x_1, x_2, \dots, x_k$ không đổi

β_2: hệ số góc của biến Y theo biến x_2 giữ các biến $x_1, x_3, \dots, x_k$ không đổi

…

β_k: hệ số góc của biến Y theo biến x_k giữ các biến $x_1, x_2, \dots, x_{k-1}$ không đổi

b. Tối ưu hóa các hệ số hồi quy của mô hình

Mục tiêu của mô hình hồi quy logistic là tối ưu hóa các hệ số hồi quy $\beta_0, \beta_1, \beta_2, \dots \beta_k$. Một trong những thuật toán phổ biến là Gradient Descent hoặc các biến thể của nó. Dưới đây là một số bước thường được thực hiện để tối ưu hóa các hệ số hồi quy:

Xác định hàm Loss:

Chọn một hàm Loss phù hợp để đánh giá hiệu suất của mô hình trên dữ liệu huấn luyện. Trong trường hợp hồi quy logistic, thường sử dụng hàm Cross-Entropy Loss. Được biểu diễn như sau:

$$Log\ Loss = J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

Trong đó:

θ là vecto hệ số hồi quy

N là số lượng mẫu trong tập dữ liệu.

y_i là giá trị thực tế của mẫu thứ i (0 hoặc 1)

$\hat{y}_i$ là xác suất dự đoán của mẫu thứ i thu được từ mô hình hồi quy logistic.

Xác định Gradient của hàm Loss:

Gradient của hàm mất mát là vector chứa các đạo hàm riêng của hàm mất mát đối với mỗi tham số. Trong trường hợp của hồi quy logistic, gradient có thể được tính toán như sau:

$$\nabla J(\theta) = \frac{1}{N} X^T(h_\theta(x) - y)$$

Trong đó:

X là ma trận dữ liệu

$h_\theta(x)$ là giả thuyết cho bài toán lặp

y là giá trị thực tế của biến phụ thuộc

Cập nhật tham số:

Sử dụng gradient tính được để cập nhật các hệ số của mô hình theo hướng giảm độ dốc của hàm mất mát. Công thức cập nhật có thể được biểu diễn như sau:

$$\theta_{i+1} = \theta_i - \alpha\nabla J(\theta_i)$$

Trong đó:

θ_i là vecto hệ số hồi quy tại bước lặp thứ i

α là tỷ số học

$\nabla J(\theta_i)$ là gradient của hàm mất mát tại θ_i

Lặp lại:

Lặp lại quá trình cập nhật tham số cho đến khi đạt được một điều kiện dừng, chẳng hạn như số lượng vòng lặp tối đa hoặc khi độ chệch của tham số không đổi đủ nhỏ, sự thay đổi của hàm mất mát dưới một ngưỡng nhất định, hoặc đạt được độ chính xác mong muốn.

3. Tiền xử lí dữ liệu

3.1. Đọc dữ liệu

Gán file "potability.csv' bằng "wp" sau đó đọc dữ liệu bằng lệnh "read.csv". Sau đó xem 10 dòng đầu tiên của dữ liệu.

```
> wp<-read.csv("~/water_potability.csv")
> head(wp,10)
         ph Hardness   Solids Chloramines  Sulfate Conductivity Organic_carbon Trihalomethanes Turbidity Potability
1        NA 204.8905 20791.32    7.300212 368.5164      564.3087      10.379783        86.99097  2.963135          0
2  3.716080 129.4229 18630.06    6.635246       NA      592.8854      15.180013        56.32908  4.500656          0
3  8.099124 224.2363 19909.54    9.275884       NA      418.6062      16.868637        66.42009  3.055934          0
4  8.316766 214.3734 22018.42    8.059332 356.8861      363.2665      18.436524       100.34167  4.628771          0
5  9.092223 181.1015 17978.99    6.546600 310.1357      398.4108      11.558279        31.99799  4.075075          0
6  5.584087 188.3133 28748.69    7.544869 326.6784      280.4679       8.399735        54.91786  2.559708          0
7 10.223862 248.0717 28749.72    7.513408 393.6634      283.6516      13.789695        84.60356  2.672989          0
8  8.635849 203.3615 13672.09    4.563009 303.3098      474.6076      12.363817        62.79831  4.401425          0
9        NA 118.9886 14285.58    7.804174 268.6469      389.3756      12.706049        53.92885  3.595017          0
10 11.180284 227.2315 25484.51    9.077200 404.0416      563.8855      17.927806        71.97660  4.370562          0
```

Hình 3.1: Code R và kết quả khi đọc 10 dòng đầu tiên của dữ liệu

3.2. Làm sạch dữ liệu

Kiểm tra các dữ liệu khuyết trong **wp.**

```
> #Số gia trị và tỉ lệ NA
> wp %>% gather(key = "variable", value = "value") %>%
+   group_by(variable) %>% summarise(na_num = sum(is.na(value)),na_per=mean(is.na(value)))

  variable        na_num na_per
  <chr>            <int>  <dbl>
1 Chloramines          0 0
2 Conductivity         0 0
3 Hardness             0 0
4 Organic_carbon       0 0
5 Potability           0 0
6 Solids               0 0
7 Sulfate            781 0.238
8 Trihalomethanes    162 0.0495
9 Turbidity            0 0
10 ph                491 0.150
```

*Hình 3.2.1: Code R và kết quả khi kiểm tra các dữ liệu khuyết trong **wp**.*

Nhận xét: Sau khi kiểm tra các dữ liệu trong wp nhận thấy có: 781 (23.8%) dữ liệu bị khuyết ở biến Sulfate, 162 (5%) dữ liệu bị khuyết ở biến Trihalomethanes và 491 (15%) ở biến pH. Vậy ta cần xử lí các dữ liệu khuyết đó.

Phương pháp được sử dụng là sử dụng các giá trị trung bình ở các biến có giá trị bị khuyết để thay thế cho các giá trị bị khuyết trong wp.

```
> #thay các giá trị khuyết bằng giá trị trung bình
> wp$ph[is.na(wp$ph)]=mean(wp$ph,na.rm=T)
>
> wp$Sulfate[is.na(wp$Sulfate)]=mean(wp$Sulfate,na.rm=T)
>
> wp$Trihalomethanes[is.na(wp$Trihalomethanes)]=mean(wp$Trihalomethanes,na.rm=T)
```

Hình 3.2.2: Code R dùng để thay các giá trị khuyết bằng giá trị trung bình

Sau khi sử dụng phương pháp thay thế bằng giá trị trung bình, ta kiểm tra lại các giá trị trong wp.

```
> #Kiểm tra lại các giá trị
> wp %>% gather(key = "variable", value = "value") %>%
+   group_by(variable) %>% summarise(na_num = sum(is.na(value)),na_per=mean(is.na(value)))
# A tibble: 10 x 3
   variable        na_num na_per
   <chr>            <int>  <dbl>
 1 Chloramines          0      0
 2 Conductivity         0      0
 3 Hardness             0      0
 4 Organic_carbon       0      0
 5 Potability           0      0
 6 Solids               0      0
 7 Sulfate              0      0
 8 Trihalomethanes      0      0
 9 Turbidity            0      0
10 ph                   0      0
```

Hình 3.2.3: Kiểm tra lại các giá trị khuyết

Nhận xét: Ta nhận thấy không còn giá trị khuyết sau khi thực hiện xử lý dữ liệu.

Tiếp theo, ta cần phân biệt được các biến liên tục và biến phân loại. Ở đây chúng ta có một biến phân loại "Potability" và còn lại là biến liên tục.

Ta nhận thấy các giá trị ở biến "Potability" bao gồm "0" và "1". Mặc dù ở đây ta có thể nhận biết chúng là biến phân loại khi nhìn qua, nhưng khi chúng ta thực hiện các thao tác ở thống kê mô tả và thống kê suy diễn, chúng ta cần xử lí biến "Potability" sao cho các lệnh sau có thể nhận biết đó là biến phân loại.

Để xử lí biến phân loại "Potability" chúng ta sử dụng lệnh "as.factor()".

```
> #Xử lí biến phân loại "Potability"
> as.factor(wp$Potability)
   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [223] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1
 [260] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [297] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [334] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [371] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [408] 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [445] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [482] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [519] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [556] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [593] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [630] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1
 [667] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [704] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [741] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [778] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [815] 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [852] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [889] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [926] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [963] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[1000] 0
 [ reached getOption("max.print") -- omitted 2276 entries ]
Levels: 0 1
```

Hình 3.2.4: Code R và kết quả sau khi xử lí biến phân loại "Potability"

3.3. Chia dữ liệu

Dữ liệu được chia thành tập huấn luyện (train_data) và tập thử nghiệm (test_data) với tỉ lệ 8:2.

```
y <- wp$Potability
set.seed(42, sample.kind = "default")
test_index <- createDataPartition(y=wp$Potability, times=1, p=0.2, list=F)
test_data <- wp[test_index,]
train_dada <- wp[-test_index,]
```

4. Thống kê mô tả

4.1. Thống kê dạng bảng

- **Thống kê số liệu mô tả cho các biến liên tục**

```
> continuous <- function(x) {c(summary(x),sd=sd(x))}
> apply(wp[,1:9],2,continuous)
               ph  Hardness     Solids Chloramines   Sulfate
Min.     0.000000  47.43200   320.9426    0.352000 129.00000
1st Qu.  6.277673 176.85054 15666.6903    6.127421 317.09464
Median   7.080795 196.96763 20927.8336    7.130299 333.77578
Mean     7.080795 196.36950 22014.0925    7.122277 333.77578
3rd Qu.  7.870050 216.66746 27332.7621    8.114887 350.38576
Max.    14.000000 323.12400 61227.1960   13.127000 481.03064
sd       1.469956  32.87976  8768.5708    1.583085  36.14261
        Conductivity Organic_carbon Trihalomethanes Turbidity
Min.       181.48375       2.200000         0.73800 1.4500000
1st Qu.    365.73441      12.065801        56.64766 3.4397109
Median     421.88497      14.218338        66.39629 3.9550276
Mean       426.20511      14.284970        66.39629 3.9667862
3rd Qu.    481.79230      16.557652        76.66661 4.5003198
Max.       753.34262      28.300000       124.00000 6.7390000
sd          80.82406       3.308162        15.76988 0.7803824
```

- **Tạo bảng tần suất cho biến phân loại**

```
> table(wp$Potability, dnn="Potability")
Potability
   0    1
1998 1278
```

4.2. Thống kê bằng đồ thị

4.2.1. Đồ thị boxplot

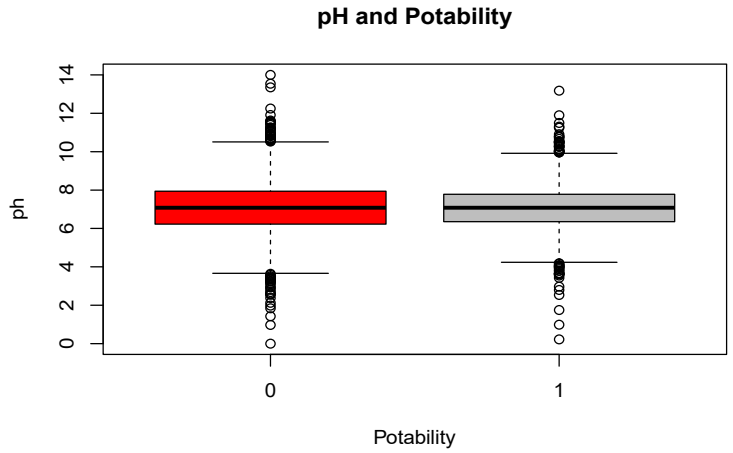
Dùng hàm boxplot() để vẽ phân phối của các biến cho từng nhóm phân loại. Xác định được giá trị lớn nhất, nhỏ nhất, tứ phân vị thứ nhất, trung vị và tứ phân vị thứ ba theo các biến.

Code R:

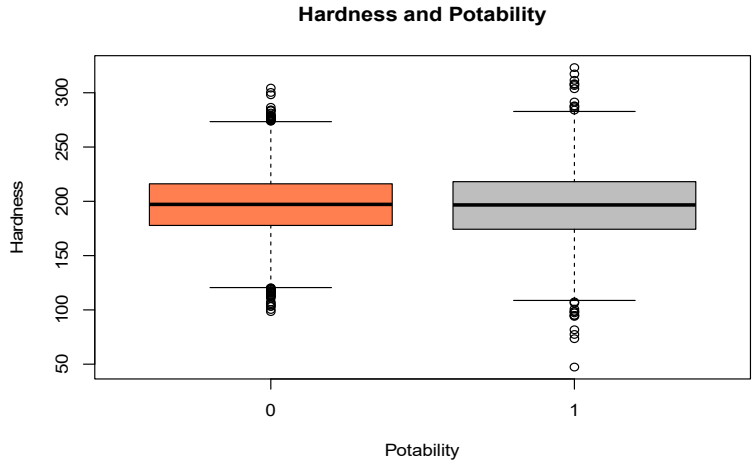
```
> boxplot(wp$ph~wp$Potability,col=c("red","gray"),xlab="Potability",ylab="ph",main="pH and Potability")
> boxplot(wp$Hardness~wp$Potability,col=c("coral","gray"),xlab="Potability",ylab="Hardness",main="Hardness and Potability")
> boxplot(wp$Solids~wp$Potability,col=c("orange","gray"),xlab="Potability",ylab="Solids",main="Solids and Potability")
> boxplot(wp$Chloramines~wp$Potability,col=c("yellow","gray"),xlab="Potability",ylab="Chloramines",main="Chloramines and Potability")
> boxplot(wp$Sulfate~wp$Potability,col=c("green","gray"),xlab="Potability",ylab="Sulfate",main="Sulfate and Potability")
> boxplot(wp$Conductivity~wp$Potability,col=c("lightblue","gray"),xlab="Potability",ylab="Conductivity",main="Conductivity and Potability")
> boxplot(wp$Organic_carbon~wp$Potability,col=c("blue","gray"),xlab="Potability",ylab="Organic_carbon",main="Organic_carbon and Potability")
> boxplot(wp$Trihalomethanes~wp$Potability,col=c("violet","gray"),xlab="Potability",ylab="Trihalomethanes",main="Trihalomethanes and Potability")
> boxplot(wp$Turbidity~wp$Potability,col=c("pink","gray"),xlab="Potability",ylab="Turbidity",main="Turbidity and Potability")
```

Kết quả thu được:

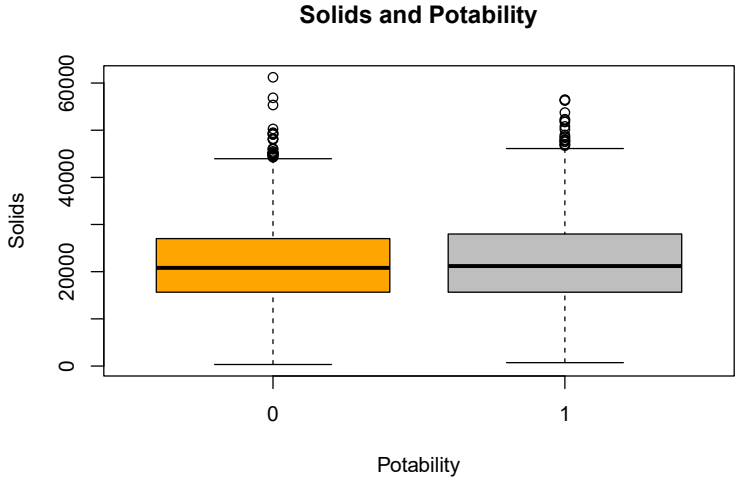
- **pH of water**



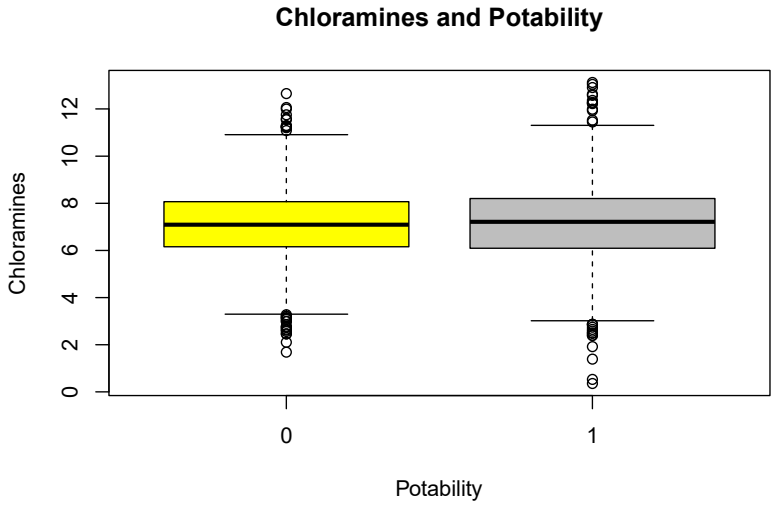
- **Hardness of water**



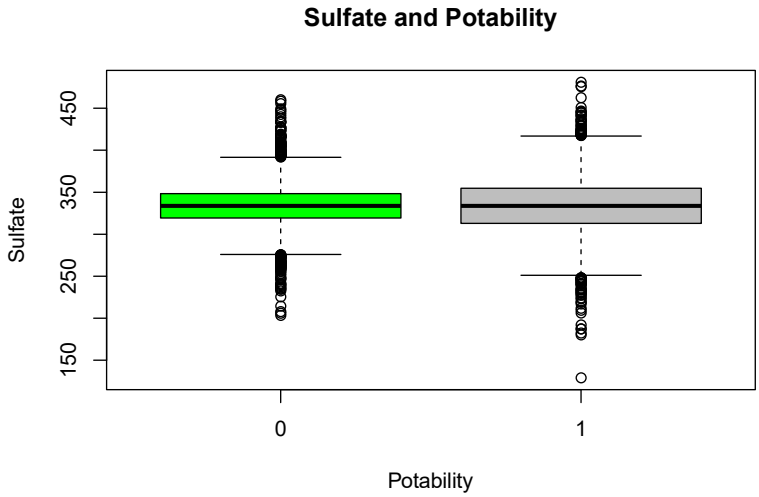
- **Solids of water**



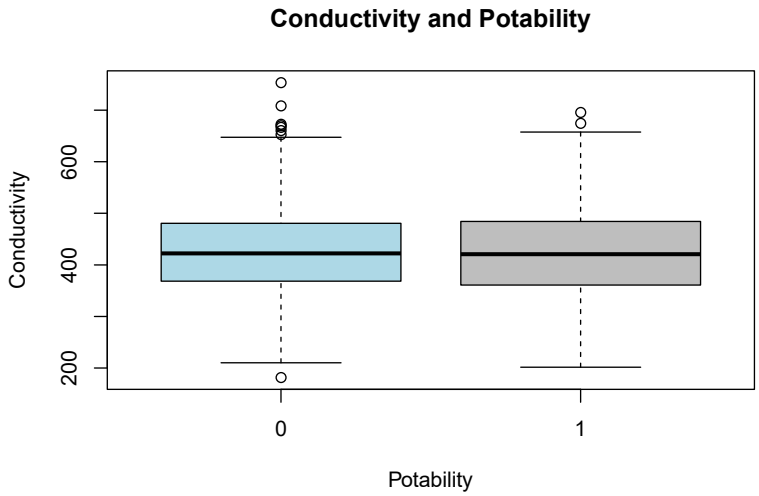
- **Chloramines of water**



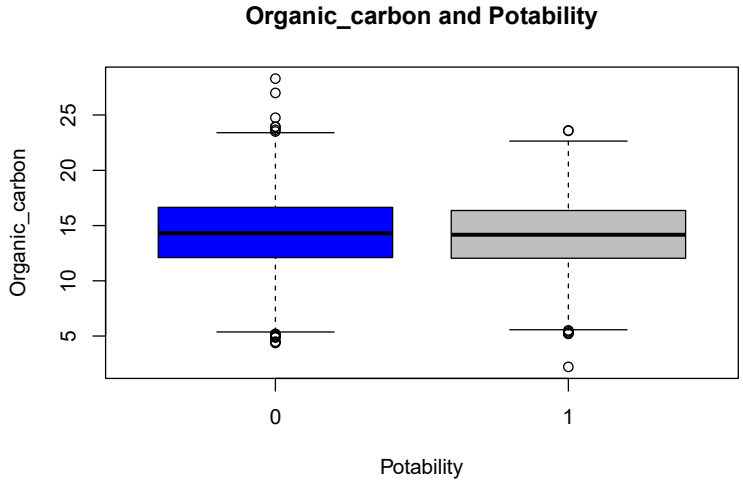
- **Sulfate of water**



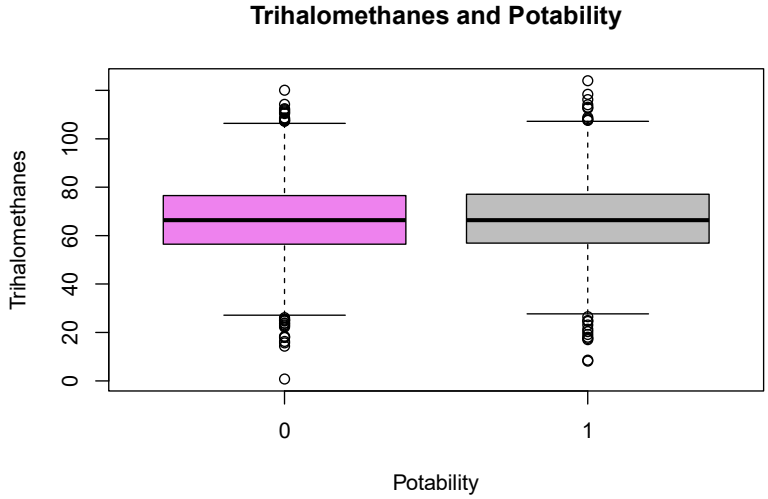
- **Conductivity of water**



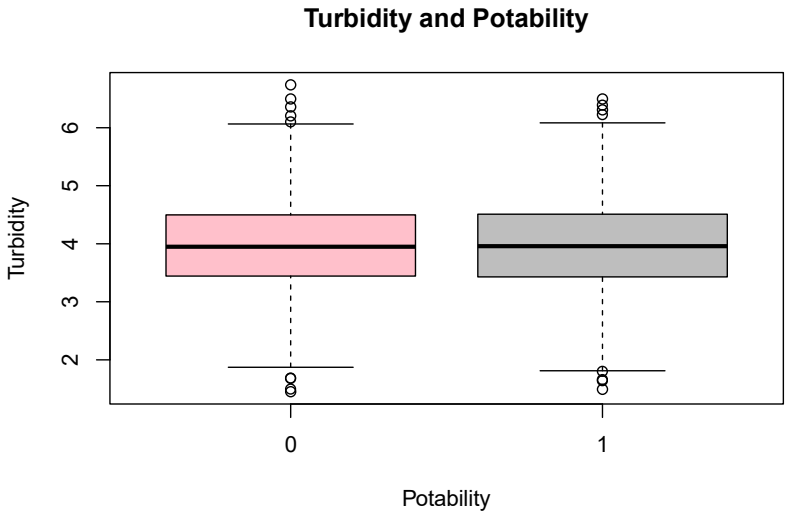
- **Organic_carbon of water**



- **Trihalomethanes of water**



- **Turbidity of water**



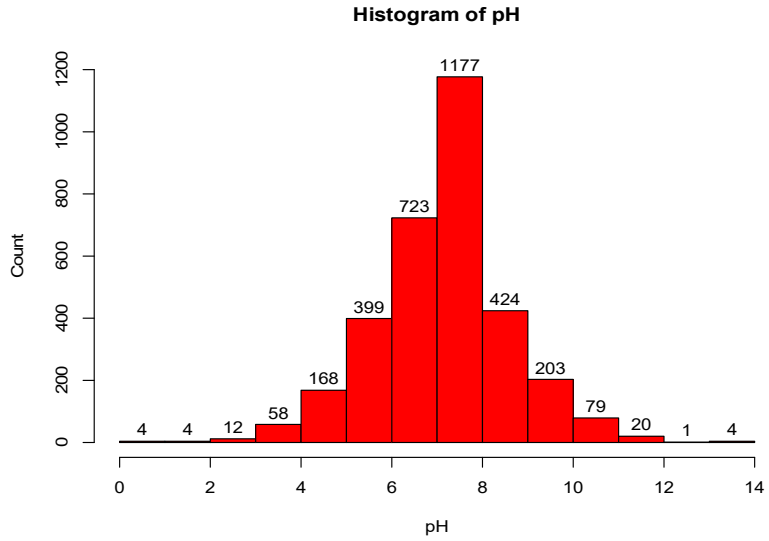
4.2.2. Đồ thị Histogram

Code R:

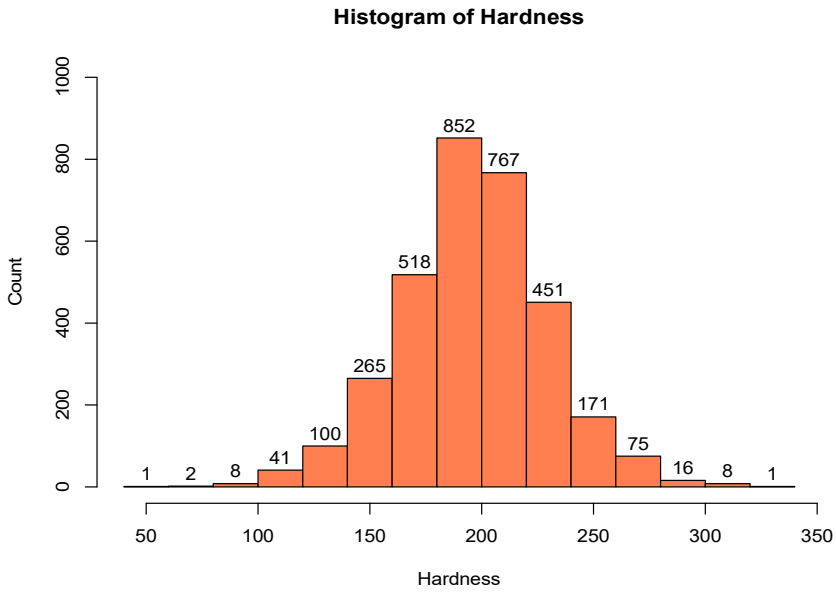
```
> hist(wp$ph,main="Histogram of pH",col="red",xlab="pH",ylab="Count",ylim=c(0,1200),labels=T)
> hist(wp$Hardness,main="Histogram of Hardness",col="coral",xlab="Hardness",ylab="Count",ylim=c(0,1000),labels=T)
> hist(wp$Solids,main="Histogram of Solids",col="orange",xlab="Solids",ylab="Count",ylim=c(0,800),labels=T)
> hist(wp$Chloramines,main="Histogram of Chloramines",col="yellow",xlab="Chloramines",ylab="Count",ylim=c(0,1000),labels=T)
> hist(wp$Sulfate,main="Histogram of Sulfate",col="green",xlab="Sulfate",ylab="Count",ylim=c(0,1400),labels=T)
> hist(wp$Conductivity,main="Histogram of Conductivity",col="lightblue",xlab="Conductivity",ylab="Count",ylim=c(0,800),labels=T)
> hist(wp$Organic_carbon,main="Histogram of Organic_carbon",col="blue",xlab="Organic_carbon",ylab="Count",ylim=c(0,800),labels=T)
> hist(wp$Trihalomethanes,main="Histogram of Trihalomethanes",col="violet",xlab="Trihalomethanes",ylab="Count",ylim=c(0,1000),labels=T)
> hist(wp$Turbidity,main="Histogram of Turbidity",col="pink",xlab="Turbidity",ylab="Count",ylim=c(0,1000),labels=T)
```

Kết quả thu được:

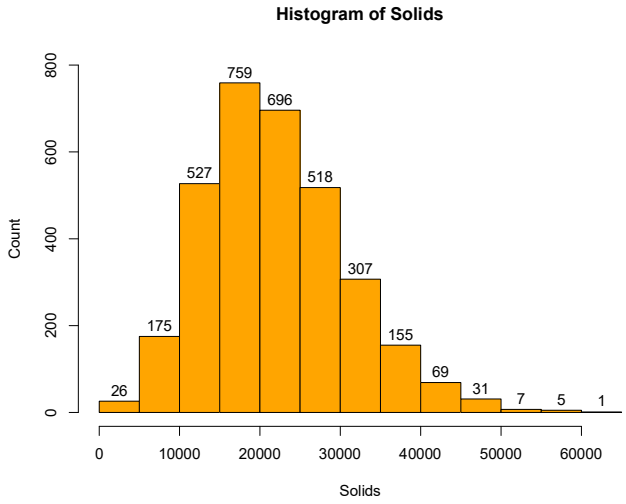
- **pH of water**



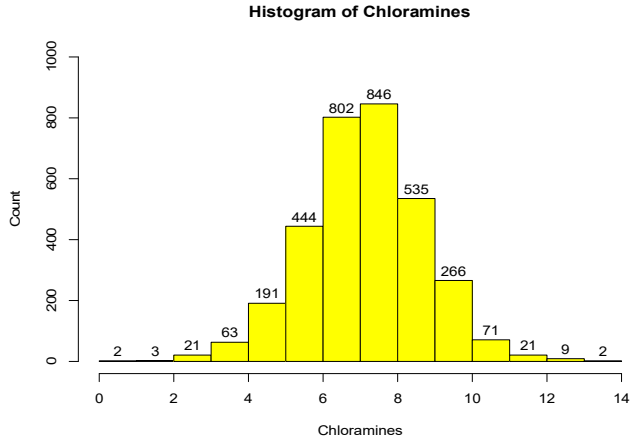
- **Hardness of water**



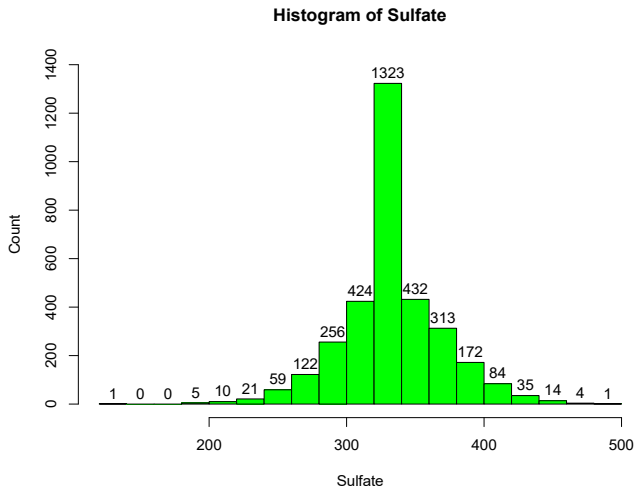
- **Chloramines of water**



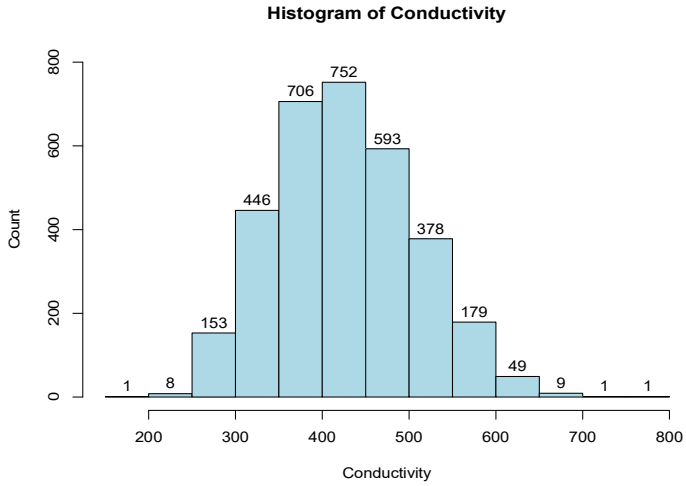
- **Chloramines of water**



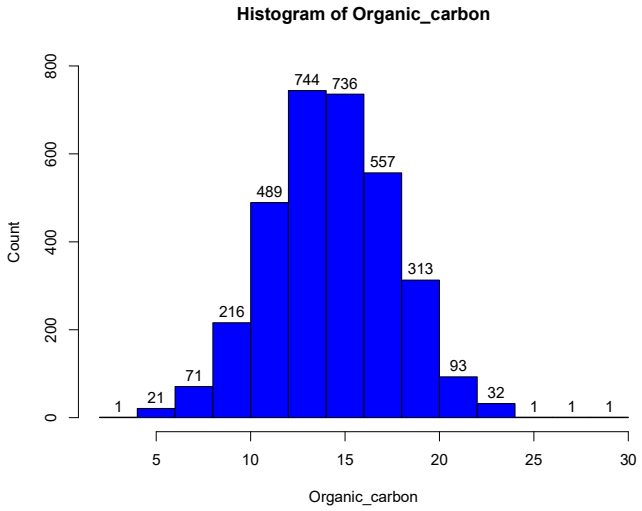
- **Sulfate of water**



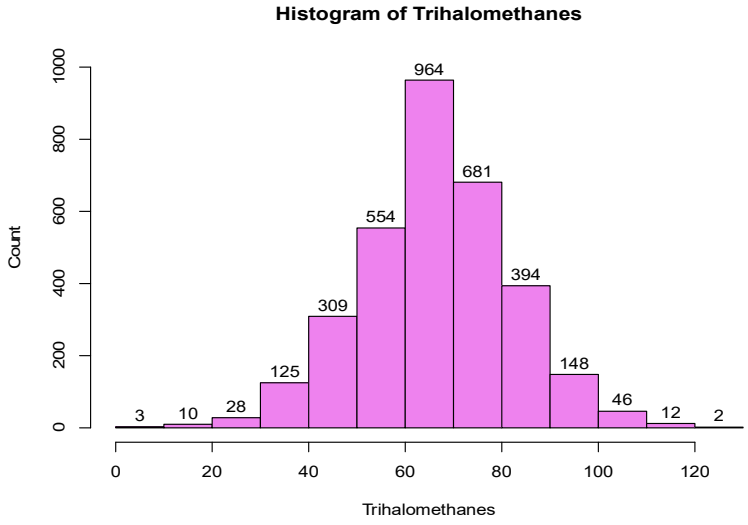
- **Conductivity of water**



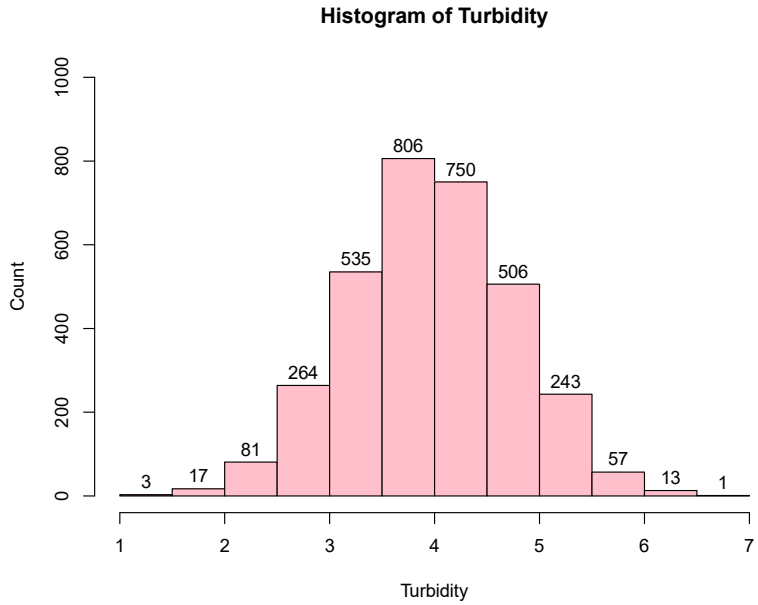
- **Organic_carbon of water**



- **Trihalomethanes of water**



- **Turbidity of water**



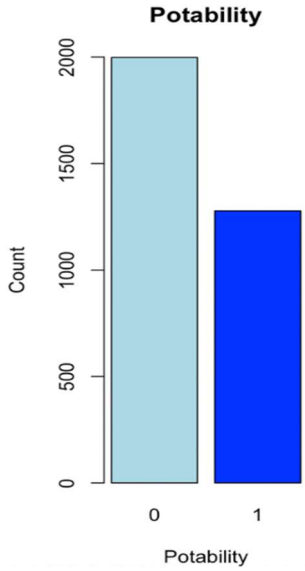
4.2.3. Đồ thị Barplot

- **Potability**

Code R:

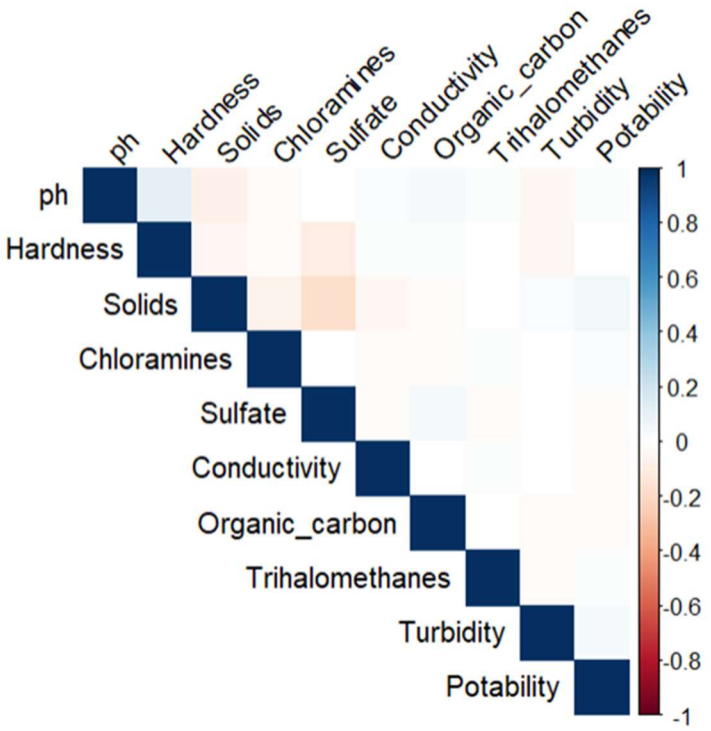
```
> barplot(table(wp$Potability),col=c("lightblue","blue"),xlab="Potability", ylab="Count",ylim=c(0,2000),main="Potability")
```

Kết quả thu được:



4.3. Kiểm tra tương quan giữa các biến

```
> y <- wp$Potability
> set.seed(42, sample.kind = "default")
> test_index <- createDataPartition(y, times=1, p=0.2, list=F)
> train_data <- wp[-test_index,]
> test_data <- wp[test_index,]
> cor_mat <- cor(train_data, use = "complete.obs")
> library(corrplot)
> corrplot(cor_mat, type = "upper",
+          tl.col = "black", tl.srt = 45, method="color")
```



Hình 4.3: Biểu đồ tương quan

Nhận xét: Từ biểu đồ tương quan, chúng ta có thể thấy rằng các biến không có mối tương quan chặt chẽ với nhau, ngoại trừ khi chúng được so sánh với chính mình. Do Potability là một biến phân loại danh nghĩa, có nghĩa là nó chỉ có thể nhận một trong hai giá trị là có thể uống được hoặc không thể uống được. Do đó chúng ta sẽ không xem xét các kết hợp thuộc tính để huấn luyện thuật toán Machine Learning.

5. Thống kê suy diễn

Trong phần này, chúng ta sẽ xây dựng các mô hình với các biến Potability, pH value, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes và Turbidity. Mục tiêu là tìm hiểu mối quan hệ giữa Potability và các biến độc lập và phát triển mô hình dự đoán ước tính chính xác Potability dựa trên các yếu tố này. Phân tích này sẽ cung cấp những hiểu biết về các yếu tố tác động đến chất lượng nguồn nước ở các vùng khác nhau và ảnh hưởng như thế nào đến khả năng uống nước.

5.1. Mô hình hồi quy Logistic

Áp dụng mô hình trên tập dữ liệu train_data và kết quả:

```
Call:
glm(formula = Potability ~ ph + Hardness + Solids + Chloramines +
    Sulfate + Conductivity + Organic_carbon + Trihalomethanes +
    Turbidity, data = train_data)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.320e-01  2.002e-01   1.159   0.2467
ph               7.389e-03  7.847e-03   0.942   0.3465
Hardness         6.622e-05  3.783e-04   0.175   0.8611
Solids           2.901e-06  1.435e-06   2.022   0.0434 *
Chloramines      8.702e-03  7.697e-03   1.131   0.2584
Sulfate         -3.053e-05  3.027e-04  -0.101   0.9197
Conductivity    -1.349e-04  1.506e-04  -0.896   0.3706
Organic_carbon  -3.550e-03  3.700e-03  -0.959   0.3375
Trihalomethanes  3.607e-04  7.533e-04   0.479   0.6321
Turbidity        1.919e-02  1.572e-02   1.220   0.2225
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2409482)

    Null deviance: 387.61  on 1608  degrees of freedom
Residual deviance: 385.28  on 1599  degrees of freedom
  (1011 observations deleted due to missingness)
AIC: 2288.2

Number of Fisher Scoring iterations: 2
```

5.2. Dự đoán độ chính xác của mô hình hồi quy Logistic

Kết quả độ chính xác của mô hình trên tập dữ liệu train_data:

```
> set.seed(42, sample.kind = "default")
> train_glm <- train(factor(Potability) ~., method = "glm",
+                    data = train_data_1)
> Accuracy (Cross-Validation)
Error: unexpected input in "Accuracy (Cross-"
> set.seed(42, sample.kind = "default")
> train_glm <- train(factor(Potability) ~., method = "glm",
+                    data = train_data_1)
> pred_glm <- predict(train_glm, test_data_1)
> confusionMatrix(pred_glm,
+                 factor(test_data_1$Potability))$overall["Accuracy"]
 Accuracy
0.6051829
```

Độ chính xác của mô hình hồi quy Logistic được đánh giá bằng cách tính toán ma trận nhầm lẫn và trích xuất giá trị độ chính xác tổng thể. Ma trận nhầm lẫn là một bảng tóm tắt hiệu suất của mô hình bằng cách so sánh các giá trị dự đoán (pred_glm) với các giá trị thực tế (factor ()). Độ chính xác tổng thể là tỷ lệ phần trăm các dự đoán chính xác được thực hiện bởi mô hình (Accuracy = 0.6051829).

5.3. Kiểm tra giả định

Dùng kiểm định Hosmer và Lemeshow để kiểm tra mức độ phù hợp của mô hình hồi quy Logistic và kết quả:

```
> hl <- hoslem.test(glm$fitted.values,fitted(glm ))
> hl

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  glm$fitted.values, fitted(glm)
X-squared = 0, df = 8, p-value = 1
```

Từ kết quả trên, giá trị p lớn (lớn hơn 0.05) cho thấy rằng chúng ta có đủ bằng chứng để kết luận sự phù hợp. Vì vậy, chúng ta chấp nhận giả thuyết không cho rằng mô hình phù hợp với dữ liệu. Do đó, chúng ta áp dụng mô hình này để dự đoán kết quả như mong đợi.

6. Thảo luận và mở rộng

6.1. Thảo luận

Với tình trạng ô nhiễm môi trường biến động mạnh mẽ như hiện nay, nước uống chắn chắc sẽ là một vấn đề nóng đối với bất kỳ một quốc gia nào. Có thể nhận thấy ở các nước kém phát triển thậm chí không thể tiếp cận với nguồn nước sạch. Do đó giải pháp tối ưu hiện nay chính là đưa ra chỉ tiêu tối thiểu cho nguồn nước: nguồn nước đảm bảo một số tiêu chí nhất định, an toàn và được cho phép dung trong thực phẩm,... từ đó đáp ứng được cơ bản nhất về vấn đề này.

Trong nhiệm vụ nghiên cứu đề tài lần nay còn giới hạn trong phạm vi tiêu chuẩn để đánh giá nguồn nước dựa vào các chỉ số vật lí, hóa học. Hơn thế, việc sử dụng thuật toán Machine Learning để tính toán, so sánh các kết quả chỉ mang tính tương đối. Xét về điều kiện thực tế, chất lượng nguồn nước là một yếu tố thay đổi liên tục theo thời gian và không đồng đều đối với từng mẫu thử. Điều quan trọng nhất có lẽ là các yếu tố hóa lí chỉ là một phần để đánh giá. Song song với đó là sự tồn tại và hoạt động của hệ vi sinh vật trong nước. Bởi vậy giả sử thuật toán đoán chính xác thì ta cũng không thể dựa vào đó để quyết định trực tiếp rằng nguồn nước có thể sử dụng được hay không. Trên thực tế, công việc này yêu cầu trang thiết bị hiện đại và quy trình phức tạp, nghiêm ngặt. Tuy nhiên việc xây dựng thuật toán nhằm phán đoán cục bộ như vậy vẫn có giá trị giúp đánh giá một cách đơn giản ban đầu để đề ra kế hoạch, lên phương án để cải tạo và sử dụng nước ở các khu vực nhất định.

Như vậy, đánh giá chất lượng nguồn nước dựa vào việc thống kê, xử lí số liệu và thuật toán phân tích các chỉ số là một biện pháp mang tính tương đối. Dù vậy nó vẫn có những ưu điểm và tồn tại như một phương pháp phổ biến và có giá trị thực tế rõ ràng đối với các nhà phân tích.

6.2. Mở rộng

Các mô hình khác nhau được huấn luyện để xem mô hình nào hoạt động tốt nhất với tập dữ liệu thử nghiệm test_data_1. Trong trường hợp này, một số thuật toán Machine Learning phân loại được sử dụng để tối ưu hóa hiệu suất của mô hình.

6.2.1. Thuật toán K – Nearest Neighbors (KNN)

Kết quả:

```
> set.seed(42, sample.kind = "default")
> train_knn <- train(factor(Potability) ~ ., method = "knn",
+                    tuneGrid = data.frame(k = seq(3, 45, 2)),
+                    data = test_data_1,
+                   trControl= train_control)
> pred_knn <- predict(train_knn, test_data_1)
> confusionMatrix(pred_knn,
+                 factor(test_data_1$Potability))$overall["Accuracy"]
 Accuracy
0.6189024
```

6.2.2. Thuật toán Naive Bayes (NB)

Kết quả:

```
> set.seed(42, sample.kind = "default")
> train_nb <- train(factor(Potability) ~., method = "naive_bayes",
+                   data = test_data_1)
> pred_nb <- predict(train_nb, test_data_1)
> confusionMatrix(pred_nb,
+                 factor(test_data_1$Potability))$overall["Accuracy"]
 Accuracy
0.6265244
```

6.2.3. Thuật toán Quantitative Discriminant Analysis (QDA)

Kết quả:

```
> set.seed(42, sample.kind = "default")
> train_qda <- train(factor(Potability) ~., data = test_data_1,
+                    method = "qda")
> pred_qda <- predict(train_qda, test_data_1)
> confusionMatrix(pred_qda,
+                 factor(test_data_1$Potability))$overall["Accuracy"]
 Accuracy
0.6768293
```

6.2.4. Thuật toán Support Vector Machine (SVM)

Kết quả:

```
> set.seed(42, sample.kind = "default")
> train_svm <- train(factor(Potability) ~., method = "lssvmRadial",
+                    data = test_data_1,
+                    tuneGrid = expand.grid(
+                      tau = c(0.001,0.009,0.01,0.1),
+                      sigma = c(0.001,0.009,0.01,0.1)),
+                    trControl = train_control)
> pred_svm <- predict(train_svm, test_data_1)
> confusionMatrix(pred_svm,
+                 factor(test_data_1$Potability))$overall["Accuracy"]
 Accuracy
0.7210366
```

Nhận xét:

Từ kết quả dự đoán của các mô hình trên, chúng ta có thể thấy rằng mô hình có độ chính xác cao nhất là mô hình SVM (Accuracy = 0.7210366), mô hình có độ chính xác thấp nhất là mô hình KNN (Accuracy = 0.6189024). Do đó, mô hình SVM là mô hình hoạt động tốt nhất.

Kết luận:

Sau khi thực hiện các bước để đạt được mục tiêu của đề tài này, bằng cách sử dụng dữ liệu của tệp The water_potability.csv cùng với thuật toán SVM được tối ưu hóa, chúng ta đã thu được kết quả tốt nhất về độ chính xác khi áp dụng mô hình để làm rõ dữ liệu. Chúng ta nhận thấy rằng thuật toán SVM rất quan trọng trong việc xác định chất lượng nước dựa trên các biến cụ thể, từ đó người dùng có thể biết được liệu nước đó có an toàn để uống hay không.

Tổng kết

Đối với đề tài "Water potability", việc nghiên cứu thực hiện các thao tác trên bảng số liệu về các chỉ số để đánh giá nguồn nước là một cơ hội để chúng em học tập, phát triển thêm các kĩ năng này. Bên cạnh đó việc yêu cầu sử dụng các phần mềm R và R studio đã giới thiệu, giúp sinh viên làm quen với một phần mềm lập trình mới, trau dồi thêm vốn kiến thức hữu ích. Việc nghiêng cứu một đề tài lớn, tích hợp nhiều nhiệm vụ: phân tích số liệu, vẽ biểu đồ, tìm kiếm thông tin về nhu cầu, yêu cầu chất lượng nguồn nước và các tiêu chí đánh giá,... và khó khăn hơn là kết hợp tất cả để viết nên thuật toán tính toán để đưa ra kết luận về chất lượng nguồn nước dựa trên các số liệu nguyên thủy nhất. Để đưa ra đáp án hoàn chỉnh, chính xác là một điều gần như bất khả thi. Chính bởi vì vậy, báo cáo dự án tất yếu sẽ có những thiếu sót, không trọn vẹn. Do đó chúng em rất hi vọng sẽ nhận được những đánh giá, góp ý chỉnh sửa từ giảng viên để có thể sửa đổi và ngày càng tốt hơn trong các dự án sắp tới.

Cuối cùng, với tư cách là nhóm nghiên cứu, chúng em hi vọng đề tài này sẽ đem lại giá trị về mặt học thuật và có thể dự đoán tương đối về điều kiện thực tế cho việc đánh giá chất lượng nguồn nước. Trân trọng và xin chân thành cảm ơn!

TÀI LIỆU THAM KHẢO

1. Douglas C. Montgomery-George C. Runger (2011), Applied Statistic and Probability for Engineers.
2. Hoàng Trọng – Chu Nguyễn Mộng Ngọc (2008), Thống kê ứng dụng trong Kinh tế - Xã hội, NXB. Thống kê.
3. https://www.analyticsvidhya.com/blog/2021/08/conceptual-understanding-of-logistic-regression-for-data-science-beginners/
4. https://s.net.vn/zeYk
5. https://s.net.vn/0Wb7
6. https://s.net.vn/7uA3
7. https://www.kaggle.com/datasets/adityakadiwal/water-potability