TRƯỜNG ĐẠI HỌC QUỐC TẾ HỒNG BÀNG

KHOA CÔNG NGHỆ THÔNG TIN

oOo

BÁO CÁO CUỐI MÔN

MÔN: KHOA HỌC DỮ LIỆU

Giảng viên hướng dẫn : Lê Văn Hạnh

Sinh viên thực hiện : Phạm Quốc Đạt

Mã số sinh viên : 2111111172

TP. Hồ Chí Minh, 2024

LỜI CẢM ƠN

Để hoàn thành đề tài này, em đã nhận được sự hướng dẫn, giúp đỡ và góp ý nhiệt tình của thầy Lê Văn Hạnh.

Em xin gửi lời biết ơn sâu sắc đến thầy Lê Văn Hạnh đã dành nhiều thời gian và tâm huyết hướng dẫn nghiên cứu và giúp em hoàn thành luận văn tốt nghiệp.

Em cũng xin chân thành cảm ơn đến quý thầy cô trường Đại học Quốc Tế Hồng Bàng, đặc biệt là những thầy cô đã tận tình dạy bảo cho em suốt thời gian học tập tại trường.

Em xin chân thành cảm ơn Ban Giám Hiệu trường Đại Học Quốc Tế Hồng Bàng cùng quý thầy cô trong Khoa Công Nghệ Thông Tin đã tạo rất nhiều điều kiện để em học tập và hoàn thành tốt khóa học.

Em đã có nhiều cố gắng hoàn thiện đề tài bằng tất cả năng lực của mình, tuy nhiên không thể tránh khỏi nhiều thiếu sót, rất mong nhận được những đóng góp quý báu của thầy và các bạn.

TRANG CAM KẾT

Em xin cam kết báo cáo này được hoàn thành dựa trên các kết quả nghiên cứu của em và các kết quả nghiên cứu này chưa được dùng cho bất cứ báo cáo cùng cấp nào khác.

TP.HCM, ngày ….. tháng ….. năm …..

Người thực hiện

Phạm Quốc Đạt

MỤC LỤC

NHẬN XÉT CỦA GIẢNG VIÊN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP.HCM, Ngày…Tháng… Năm 2024

Chữ ký giảng viên

DANH MỤC BIỂU ĐỒ HÌNH VẼ

DANH MỤC BẢNG BIỂU

CHƯƠNG 1: GIỚI THIỆU VỀ CƠ SỞ DỮ LIỆU SỬ DỤNG CHO ĐỀ TÀI

1.1. Tổng quan về cơ sở dữ liệu

Tập dataset mà em thu thập và phân tích có tên là "***Medical Cost Personal Datasets***", có ý nghĩa là ***tập dữ liệu cá nhân về chi phí y tế***. Tập dataset này được em thu tập từ nguồn Kaggle: Your Home for Data Science.

1.2. Giới thiệu các thuộc tính (fields)

Tập dataset này có 1338 records (rows) và 7 fields (columns).

1.2.1. Tên của các fields

Thứ tự	Tên các field
1	Age
2	Sex
3	Bmi
4	Children
5	Smoker
6	Region
7	Charges

Bảng 1 Tên của các fields

1.2.2. Ý nghĩa chi tiết của các fields trong tập dữ liệu

Thuộc tính (field)	Ý nghĩa
Age	Tuổi của người thụ hưởng chính
Sex	Giới tính người đấu thầu bảo hiểm (nam hoặc nữ)
Bmi	Chỉ số khối cơ thể, cung cấp sự hiểu biết về cơ thể, cân nặng tương đối cao hoặc thấp so với chiều cao, chỉ số khách quan về trọng lượng cơ thể (kg / m ^ 2) sử dụng tỷ lệ giữa chiều cao và cân nặng, lý tưởng là 18,5 đến 24,9
Children	Số trẻ em được bảo hiểm y tế / Số người phụ thuộc
Smoker	Có hút thuốc hoặc không (yes/no)

Region	Khu vực cư trú của người thụ hưởng ở Mỹ, đông bắc, đông nam, tây nam hoặc tây bắc.
Charges	Chi phí y tế cá nhân cho mỗi người

Bảng 2 Ý nghĩa các fields

1.2.3. Số giá trị null của các fields

Thuộc tính (field)	Số giá trị null
Age	0
Sex	0
Bmi	0
Children	0
Smoker	0
Region	0
Charges	0

Bảng 3 Số giá trị null của các fields

1.2.4. Số giá trị unique của các fields

Thuộc tính (field)	Số giá trị unique
Age	0
Sex	0
Bmi	0
Children	0
Smoker	0
Region	0
Charges	0

Bảng 4 Số giá trị unique của các fields

1.2.5. Kiểu dữ liệu của các field

Thuộc tính (field)	Kiểu dữ liệu
Age	Int
Sex	String

Bmi	Decimal
Children	Int
Smoker	Boolean
Region	String
Charges	Decimal

Bảng 5 Kiểu dữ liệu của các fields

1.2.5.1. Field age

AGE							
Mean	Midrange	Mode	Five-number summary				
39.2	41	18	Min	Q1	Median	Q3	Max
		Xuất hiện 69 lần	18	27	39	51	64

Bảng 6 Field age

1.2.5.2. Field bmi

BMI							
Mean	Midrange	Mode	Five-number summary				
30.7	34.545	32.3	Min	Q1	Median	Q3	Max
		Xuất hiện 13 lần	15.96	26.3	30.4	34.7	53.13

Bảng 7 Field bmi

1.2.5.3. Field children

CHILDREN							
Mean	Midrange	Mode	Five-number summary				
1.09	2.5	0	Min	Q1	Median	Q3	Max
		Xuất hiện 574 lần	0	0	1	2	5

Bảng 8 Field chidren

1.2.5.4. Field charges

CHARGES							
Mean	Midrange	Mode	Five-number summary				
13270.4223	32446.152	1639.5631	Min	Q1	Median	Q3	Max
		Xuất hiện 2 lần	1121.87	4740.28	9382.033	16639.912	63770.42

Bảng 9 Field charges

1.2.5.5. Field sex

SEX			
Giá trị	Số lượng	Chiếm (%)	Mode
male	676	50.52	✓
female	662	49.48	
Tổng	1338	100	

Bảng 10 Field sex

1.2.5.6. Field smoker

SMOKER			
Giá trị	Số lượng	Chiếm (%)	Mode
yes	274	20.48	
no	1064	79.52	✓
Tổng	1338	100	

Bảng 11 Field smoker

1.2.5.7. Field region

REGION			
Giá trị	Số lượng	Chiếm (%)	Mode
southeast	364	27.2	✓
southwest	325	24.3	
northeast	324	24.2	
northwest	325	24.3	
Tổng	1338	100	

Bảng 12 Field region

CHƯƠNG 2: PHÂN TÍCH – THỐNG KÊ TRÊN CƠ SỞ DỮ LIỆU ĐÃ CHỌN

2.1. Tìm hiểu dữ liệu

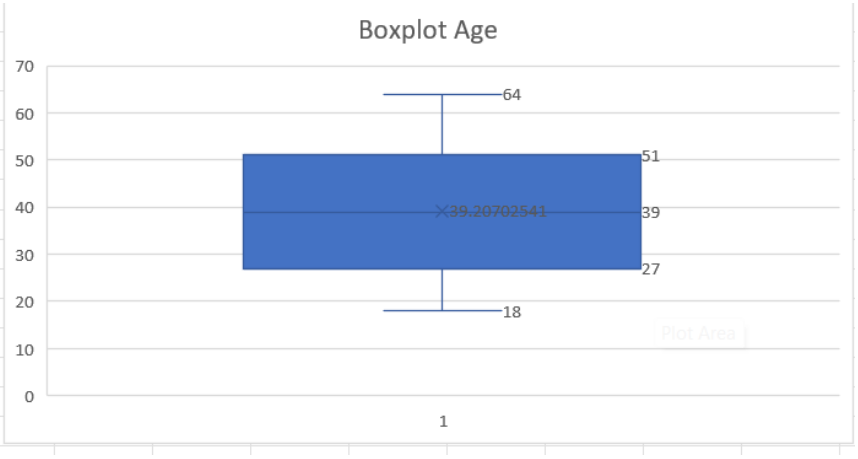
2.1.1. Chọn 3 thuộc tính để vẽ các đồ thị

3 thuộc tính em chọn để vẽ các đồ thị là: age, bmi và children

2.1.1.1. Boxplot dựa trên five-number summary

2.1.1.1.1. Age

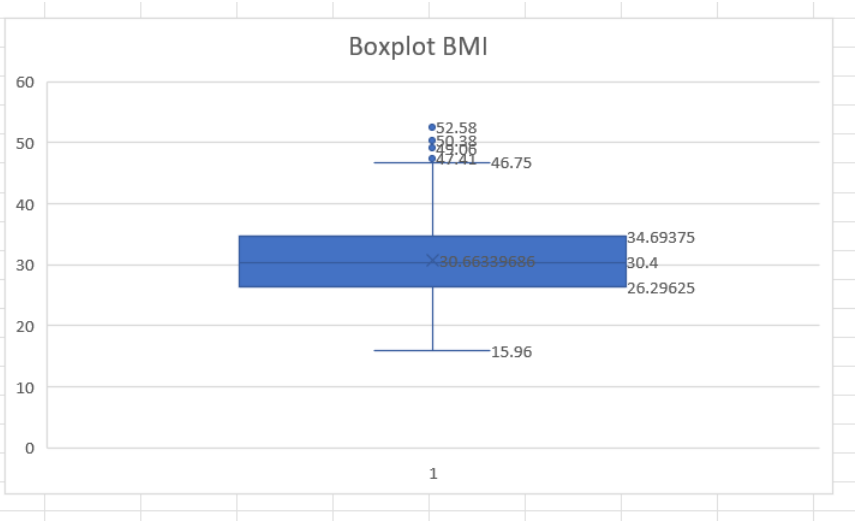
5 NUMBER SUMMARY AGE	
MIN	18
Q1	27
MEDIAN	39
Q3	51
MAX	64



Hình 2.1 Boxplot age

2.1.1.1.2. Bmi

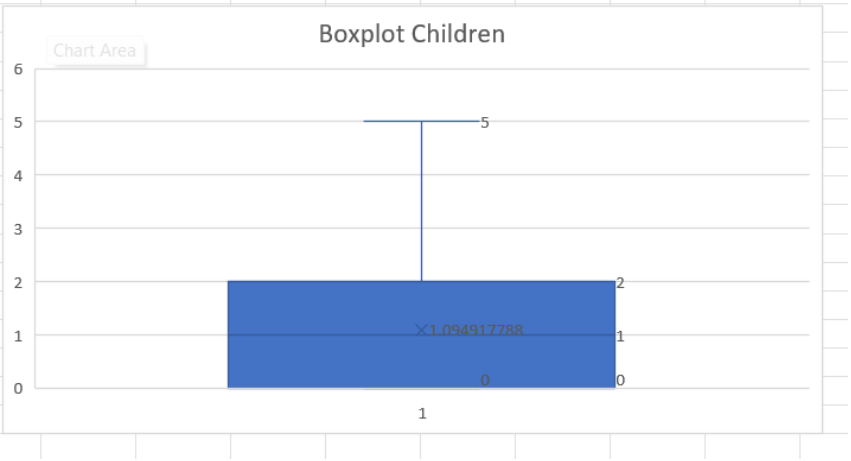
5 NUMBER SUMMARY BMI	
MIN	15.96
Q1	26.29625
MEDIAN	30.4
Q3	34.69375
MAX	53.13



Hình 2.2 Boxplot bmi

2.1.1.1.3. Children

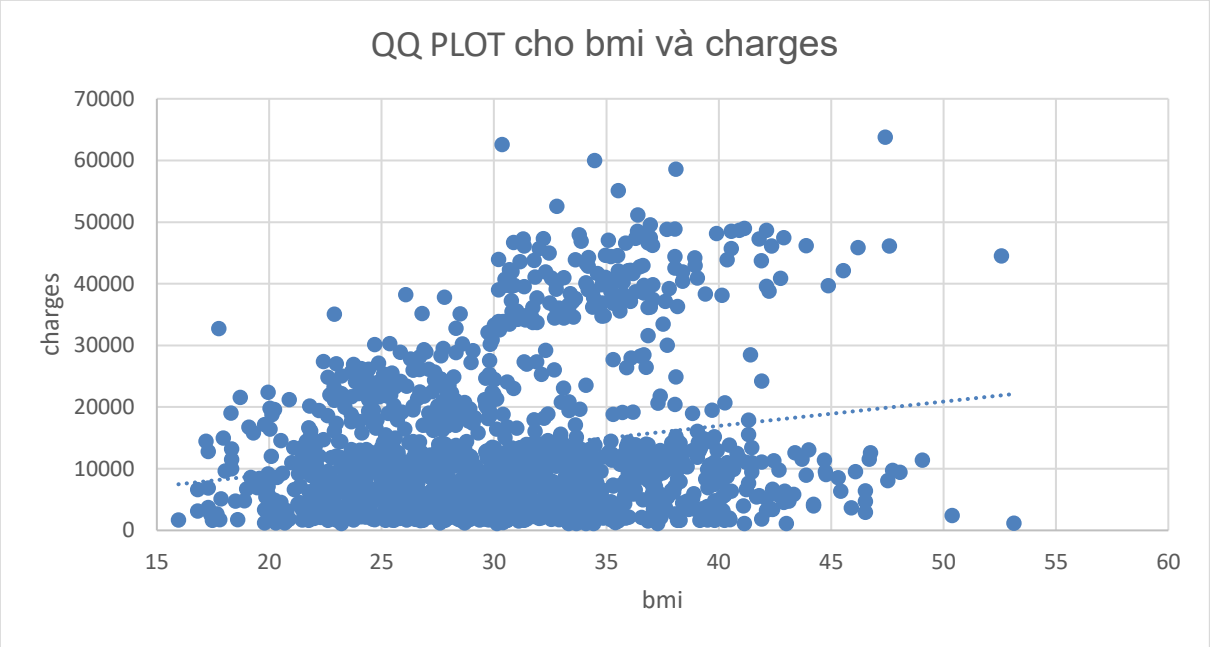
5 NUMBER SUMMARY CHILDREN	
MIN	0
Q1	0
MEDIAN	1
Q3	2
MAX	5



Hình 2.3 Boxplot children

2.1.1.2. Quantile–Quantile Plot trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

Ở đây 2 thuộc tính liên quan về ý nghĩa em chọn là: bmi và charges. Chỉ số BMI (Body Mass Index) có thể ảnh hưởng đến chi phí y tế (charges), vì những người có chỉ số BMI cao thường có nguy cơ cao hơn về nhiều vấn đề sức khỏe, điều này có thể dẫn đến các chi phí y tế cao hơn.

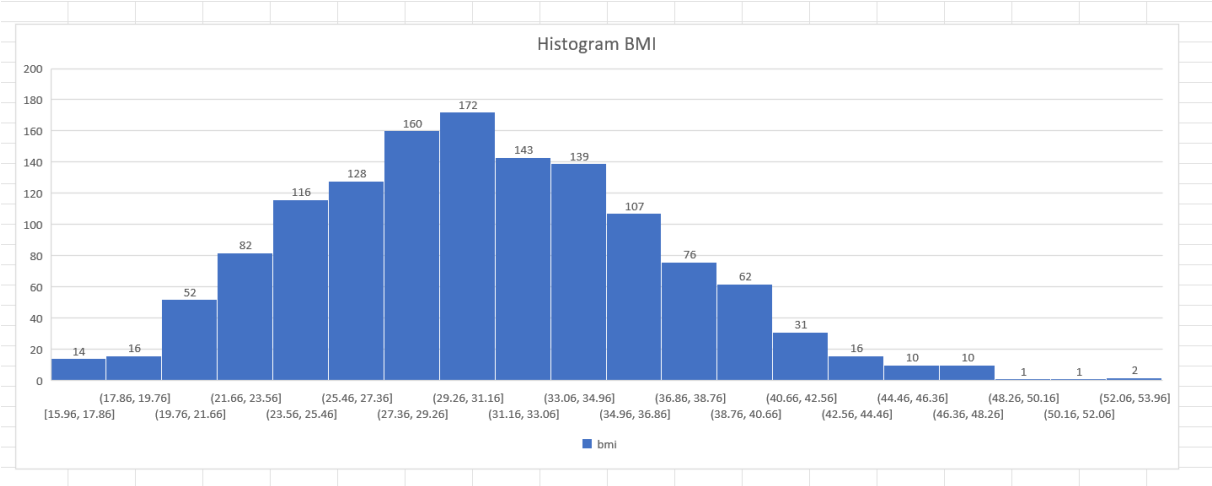


Hình 2.4 Q-Q Plot giữa bmi và charges

2.1.1.3. Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

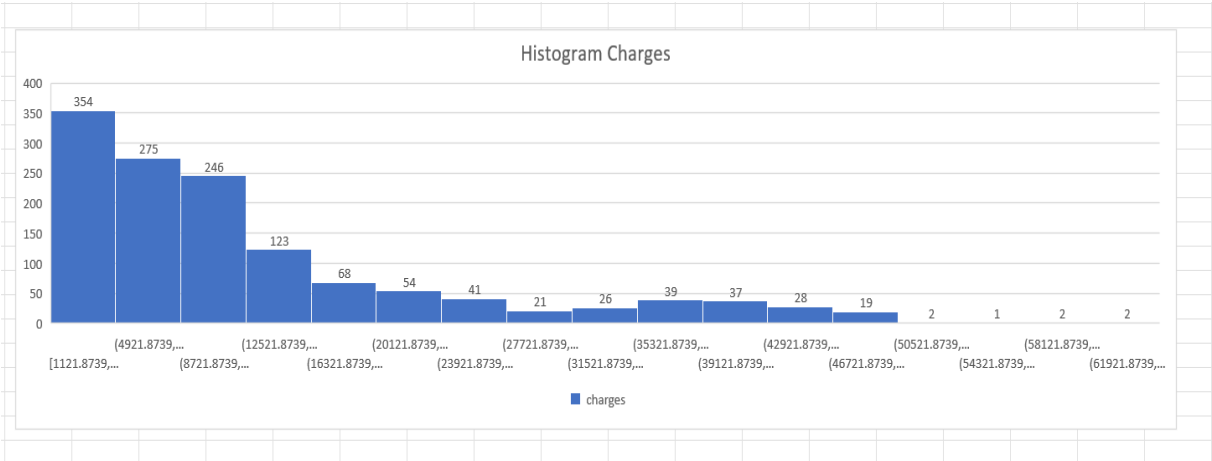
Ở đây 2 thuộc tính liên quan về ý nghĩa em chọn tiếp tục là: bmi và charges.

2.1.1.3.1. Histogram của BMI



Hình 2.5 Histogram của bmi

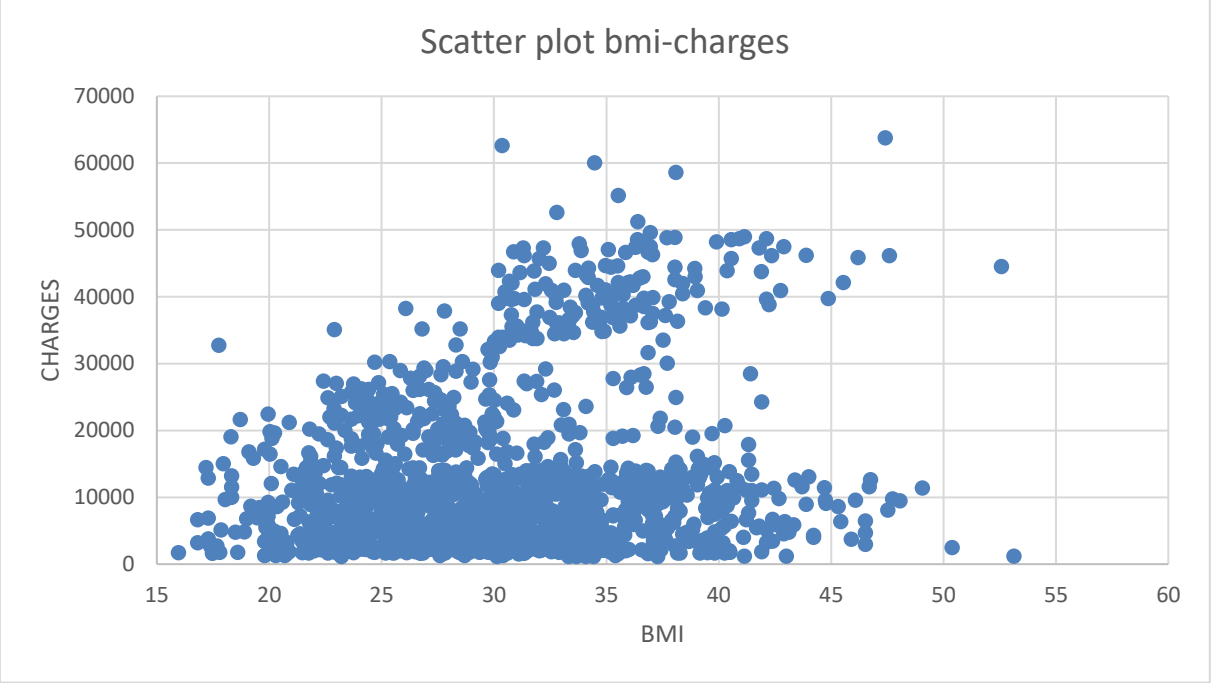
2.1.1.3.2. Histogram của Charges



Hình 2.6 Histogram của charges

2.1.1.4. Scatter trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

Ở đây 2 thuộc tính liên quan về ý nghĩa em chọn tiếp tục là: bmi và charges. Chỉ số BMI (Body Mass Index) có thể ảnh hưởng đến chi phí y tế (charges), vì những người có chỉ số BMI cao thường có nguy cơ cao hơn về nhiều vấn đề sức khỏe, điều này có thể dẫn đến các chi phí y tế cao hơn.



Hình 2.7 Scatter dựa trên 2 field là bmi và charges

2.1.2. Nhóm dữ liệu đang có theo một thuộc tính dạng danh nghĩa

Ở đây em chọn thuộc tính region gồm 4 giá trị là: southeast, southwest, northeast và northwest.

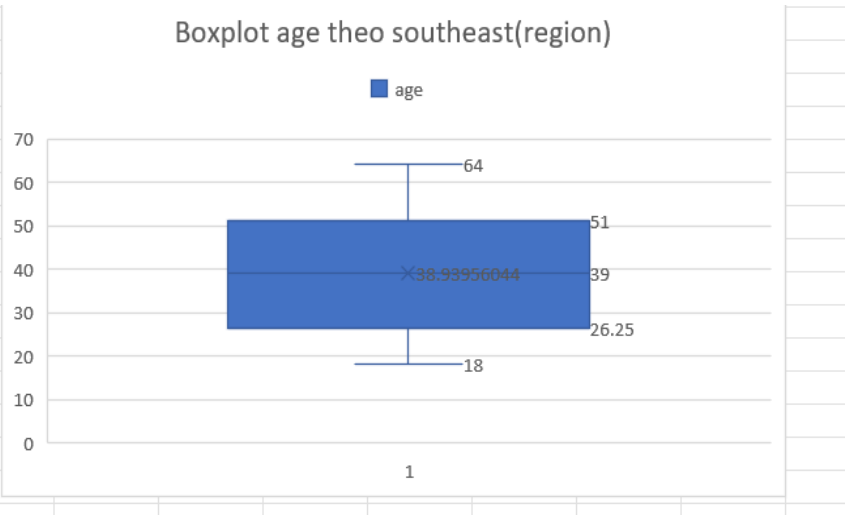
2.1.2.1. Boxplot dựa trên five-number summary

Filter theo field region gồm 4 giá trị là: southeast, southwest, northeast và northwest.

2.1.2.1.1. Filter theo southeast

Boxplot của field age sau khi đã filter theo southeast:

5 NUMBER SUMMARY - AGE	
MIN	18
Q1	26.75
MEDIAN	39
Q3	51
MAX	64

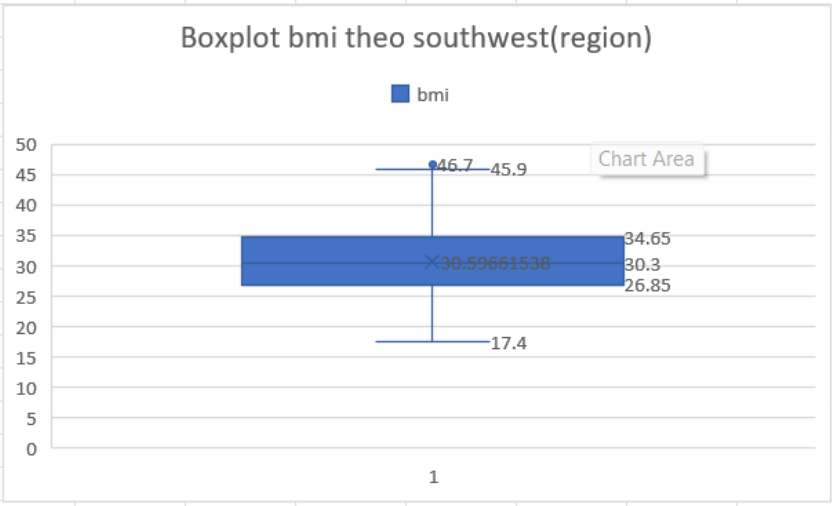


Hình 2.8 Boxplot age theo southeast

2.1.2.1.2. Filter theo southwest

Boxplot của field bmi sau khi đã filter theo southwest:

5 NUMBER SUMMARY - BMI	
MIN	17.4
Q1	26.9
MEDIAN	30.3
Q3	34.6
MAX	47.6

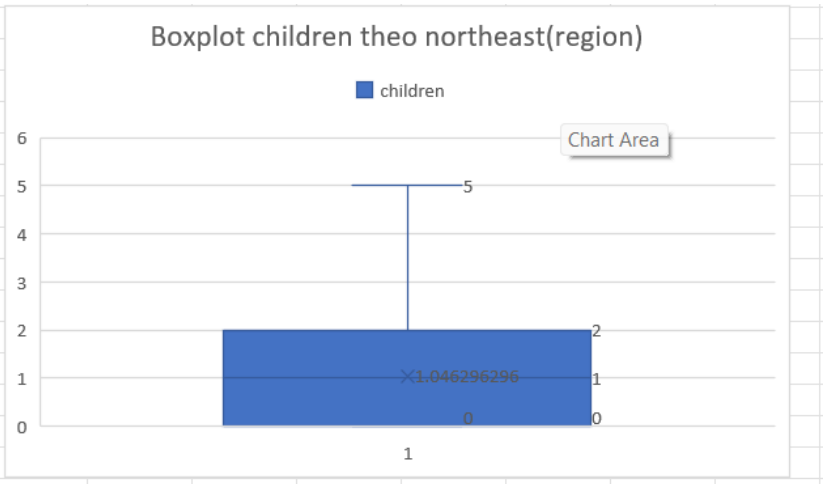


Hình 2.9 Boxplot bmi theo southwest

2.1.2.1.3. Filter theo northeast

Boxplot của field children sau khi đã filter theo northeast:

5 NUMBER SUMMARY - CHILDREN	
MIN	0
Q1	0
MEDIAN	1
Q3	2
MAX	5

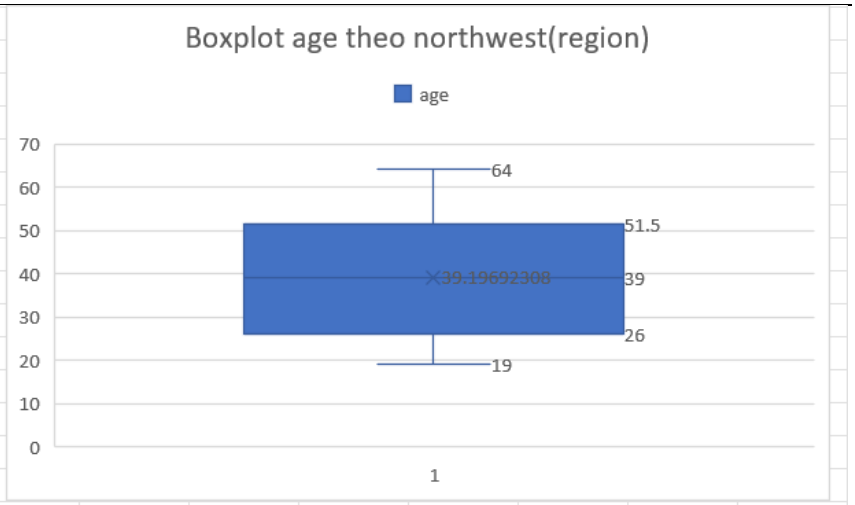


Hình 2.10 Boxplot children theo northeast

2.1.2.1.4. Filter theo northwest

Boxplot của field age sau khi đã filter theo northwest:

5 NUMBER SUMMARY - AGE	
MIN	19
Q1	26
MEDIAN	39
Q3	51
MAX	64

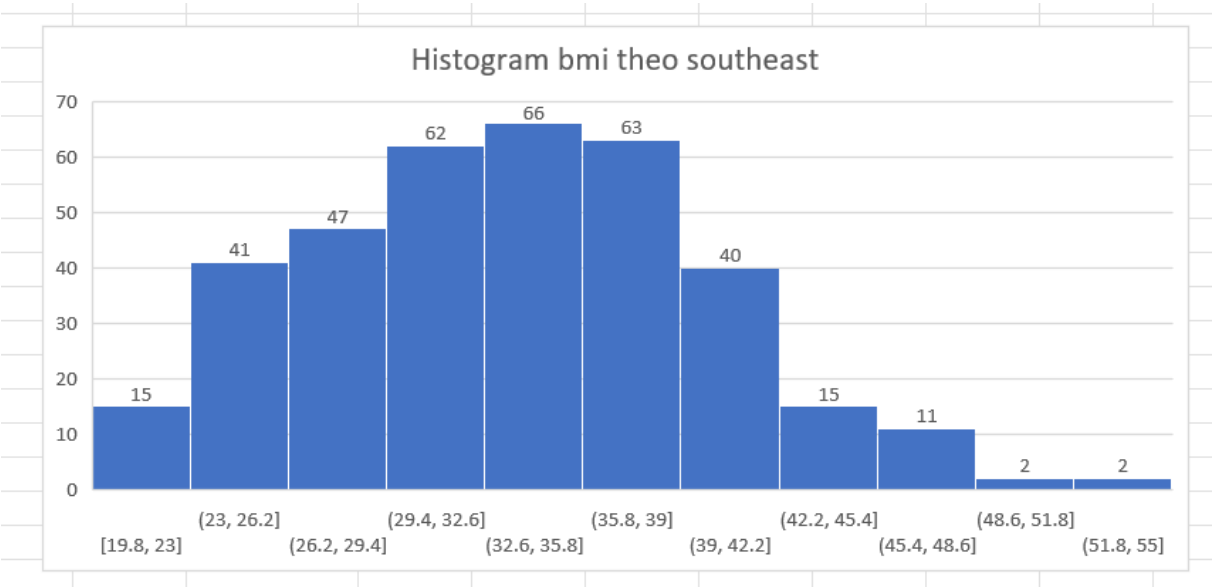


Hình 2.11 Boxplot age theo northwest

2.1.2.2. Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

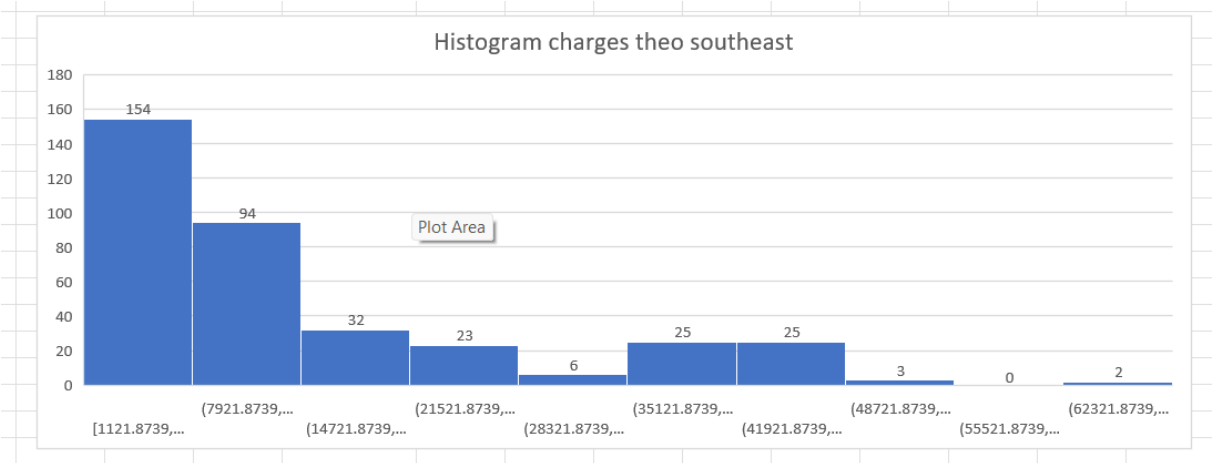
Histogram của 2 field bmi và charges sau khi filter theo southeast:

2.1.2.2.1. BMI



Hình 2.12 Histogram của bmi theo southeast

2.1.2.2.2. CHARGES:



Hình 2.13 Histogram của charges theo southeast

2.1.3. Đo lường sự tương đồng và khác biệt của dữ liệu bằng 2 cách: ma trận tương quan và độ đo Cosin

- Bốn thuộc tính dùng để vẽ ma trận tương quan và tính độ đo Cosin em chọn là: age, children, smoker và region.
- Chọn tối thiểu 4 dòng của dữ liệu đã lọc:

Thứ tự	Age	Children	Smoker	Region
A	19	0	Yes	Southwest
B	18	1	No	Southeast
C	33	0	No	Northwest
D	37	2	No	Northeast
E	28	3	No	Southeast

Bảng 13 Bảng năm dòng dữ liệu dùng để đo lường

2.1.3.1. Ma trận tương quan

Trước tiên, ta cần tính ma trận sai phân của từng thuộc tính:

2.1.3.1.1. Ma trận sai phân thuộc tính dạng danh nghĩa (region)

	A	**B**	**C**	**D**	**E**
A	0				
B	1	0			
C	1	1	0		
D	1	1	1	0	

E	1	0	1	1	0

Bảng 14 Ma trận sai phân thuộc tính danh nghĩa

2.1.3.1.2. Ma trận sai phân thuộc tính dạng nhị phân (smoker)

	A	**B**	**C**	**D**	**E**
A	0				
B	1	0			
C	1	0	0		
D	1	0	0	0	
E	1	0	0	0	0

Bảng 15 Ma trận sai phân thuộc tính nhị phân

2.1.3.1.3. Ma trận sai phân thuộc tính dạng số (age)

Sử dụng khoảng cách Manhattan:

	A	**B**	**C**	**D**	**E**
A	0				
B	1	0			
C	14	15	0		
D	18	19	4	0	
E	9	10	5	9	0

Bảng 16 Ma trận sai phân thuộc tính số

2.1.3.1.4. Ma trận sai phân thuộc tính dạng thứ tự (children)

Dữ liệu ban đầu

ID	children
A	0
B	1
C	0
D	2
E	3

→

Quy ước về thứ hạng

Children	rank
0	1
1	2
2	3
3	4

→

Xếp hạng cho dữ liệu gốc

ID	children	rank
A	0	1
B	1	2
C	0	1
D	2	3
E	3	4

Hình 2.14 Xếp hạng thuộc tính thứ tự

- Giá trị 0 (Rank =1) = $\frac{1-1}{4-1}$ = 0
- Giá trị 1 (Rank =2) = $\frac{2-1}{4-1}$ = 0,3333
- Giá trị 2 (Rank =3) = $\frac{3-1}{4-1}$ = 0,667
- Giá trị 3 (Rank =4) = $\frac{4-1}{4-1}$ = 1

Suy ra chia được 3 khoảng giá trị:

Thay thế giá trị:

ID	Children
A	0
B	0,3333
C	0
D	0,6667
E	1

Bảng 17 Thay thế khoảng giá trị

⇨ Ma trận sai phân (dùng khoảng cách Manhattan):

	A	B	C	D	E
A	0				
B	0,3333	0			
C	0	0,3333	0		
D	0,6667	0,3334	0,6667	0	
E	1	0,6667	1	0,3333	0

Bảng 18 Ma trận sai phân thuộc tính thứ tự

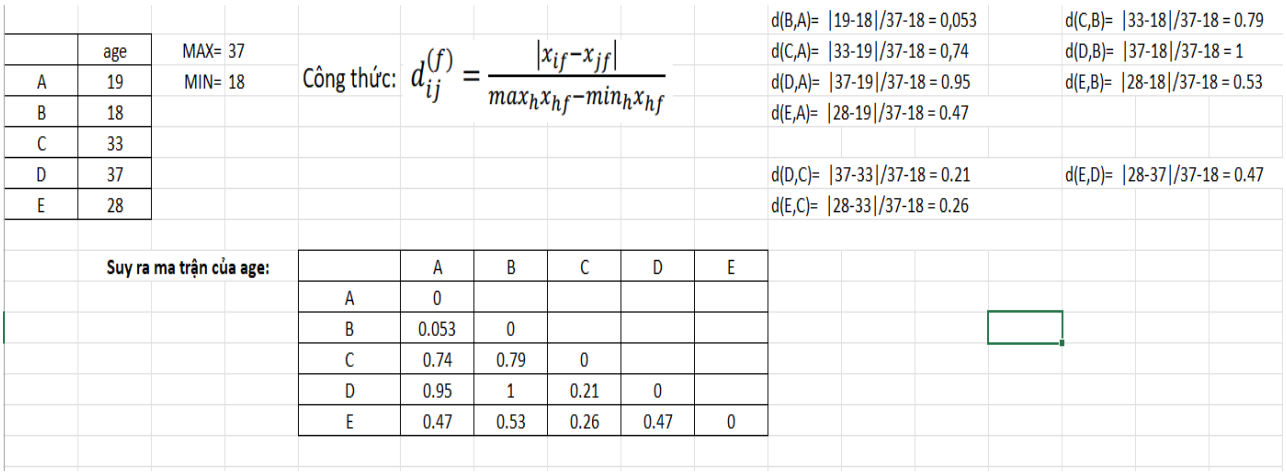
2.1.3.1.5. Ma trận tương quan (hỗn hợp)

Dựa vào các ma trận sai phân của 4 thuộc tính trên, ta có thể tính được ma trận hỗn hợp dựa vào công thức:

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

Hình 2.15 Công thức tính ma trận tương quan

Chuẩn hóa lại thuộc tính numeric (Age) về khoảng [0.0;1.0]:



Hình 2.16 Chuẩn hóa lại thuộc tính số (Age)

Dựa vào công thức trên, ta suy ra được ma trận hỗn hợp với các giá trị của ma trận như sau:

- $d(B,A) = \frac{(1x1)+(1x1)+(1x0{,}053)+(1x0{,}3333)}{1+1+1+1} = 0.6$

- d(C,A) = $\frac{(1x1)+(1x1)+(1x0{,}74)+(1x0)}{1+1+1+1} = 0.685$
- d(D,A) = $\frac{(1x1)+(1x1)+(1x0{,}95)+(1x0{,}6667)}{1+1+1+1} = 0.9$
- d(E,A) = $\frac{(1x1)+(1x1)+(1x0{,}47)+(1x1)}{1+1+1+1} = 0.88$
- d(C,B) = $\frac{(1x1)+(1x0)+(1x0{,}79)+(1x0{,}3333)}{1+1+1+1} = 0.53$
- d(D,B) = $\frac{(1x1)+(1x0)+(1x1)+(1x0{,}3334)}{1+1+1+1} = 0.58$
- d(E,B) = $\frac{(1x0)+(1x0)+(1x0{,}53)+(1x0{,}6667)}{1+1+1+1} = 0.3$
- d(D,C) = $\frac{(1x1)+(1x0)+(1x0{,}21)+(1x0{,}6667)}{1+1+1+1} = 0.5$
- d(E,C) = $\frac{(1x1)+(1x0)+(1x0{,}26)+(1x1)}{1+1+1+1} = 0.6$
- d(E,D) = $\frac{(1x1)+(1x0)+(1x0{,}47)+(1x1)}{1+1+1+1} = 0.62$

Suy ra, ta có ma trận tương quan như sau:

	A	**B**	**C**	**D**	**E**
A	0				
B	0.6	0			
C	0.685	0.53	0		
D	0.9	0.58	0.5	0	
E	0.88	0.3	0.6	0.62	0

Bảng 19 Ma trận tương quan

2.1.3.2. Độ đo Cosin

Ta cần chuẩn hóa các thuộc tính như danh nghĩa và nhị phân về dạng số, và chuyển các thuộc tính thành các vector:

Đây là dữ liệu ban đầu của 4 thuộc tính:

age	children	smoker	region
19	0	yes	southwest
18	1	no	southeast
33	0	no	northwest
37	2	no	northeast
28	3	no	southeast

2.1.3.2.1. Chuẩn hóa thuộc tính danh nghĩa và nhị phân về dạng số:

age	children	smoker	southwest	southeast	northwest	northeast
19	0	1	1	0	0	0
18	1	0	0	1	0	0
33	0	0	0	0	1	0
37	2	0	0	0	0	1
28	3	0	0	1	0	0

2.1.3.2.2. Các vecto thu được

- Vecto A = [19,0,1,1,0,0,0]
- Vecto B = [18,1,0,0,1,0,0]
- Vecto C = [33,0,0,0,0,1,0]
- Vecto D = [37,2,0,0,0,0,1]
- Vecto E = [28,3,0,0,1,0,0]

2.1.3.2.3. Thực hiện tính độ tương quan giữa các vecto:

Áp dụng công thức : sin(a,b) = $\frac{a \,.\, b}{||a|| \,.\, ||b||}$

sin(A,B) = $\frac{342}{344{,}003}$ = 0,9942

sin(A,C) = $\frac{627}{629{,}023}$ = 0,9968

sin(A,D) = $\frac{703}{706{,}231}$ = 0,9954

sin(A,E) = $\frac{532}{536{,}863}$ = 0,991

sin(B,C) = $\frac{594}{596{,}104}$ = 0,9965

sin(B,D) = $\frac{668}{669{,}271}$ = 0,9981

sin(B,E) = $\frac{508}{508{,}767}$ = 0,9985

sin(C,D) = $\frac{1221}{1223{,}789}$ = 0,9977

sin(C,E) = $\frac{924}{930{,}301}$ = 0,9932

sin(D,E) = $\frac{1042}{1044{,}488}$ = 0,9976

Kết luận

- Cặp B,E có độ tương đồng cao nhất với giá trị độ tương tự cosin = 0,9985
- Cặp A,E có độ tương đồng thấp nhất với giá trị độ tương tự cosin = 0,991

2.1.3.2.4. So sánh 2 cách đo lường:

Giống nhau

- Cả hai phương pháp đều được sử dụng để đo lường mối quan hệ hoặc tương đồng giữa các đối tượng.
- Cả hai phương pháp đều được áp dụng trong phân tích dữ liệu đa biến, trong đó có nhiều biến được quan sát và phân tích.

Khác nhau

Về phương pháp để đo lường:

- Ma trận tương quan đo lường mối quan hệ tuyến tính giữa các biến.
- Độ đo cosin đo lường sự tương đồng không gian giữa các vectơ.

Về cách tính toán:

- Ma trận tương quan thường được tính bằng cách tính toán ma trận hiệp phương sai hoặc ma trận tương quan Pearson.
- Độ đo cosin thường được tính bằng cách tính cosin của góc giữa hai vectơ.

Về phạm vi:

- Ma trận tương quan thường được sử dụng trong phân tích dữ liệu số liệu hoặc dữ liệu đa biến, đặc biệt trong việc hiểu các mối quan hệ giữa các biến và loại bỏ đồng biến tuyến tính.
- Độ đo cosin thường được sử dụng trong các bài toán gom cụm và tìm kiếm thông tin, trong đó cần đo lường sự tương đồng giữa các đối tượng.

Tóm lại, mặc dù cả hai phương pháp đều được sử dụng để đo lường mối quan hệ hoặc tương đồng giữa các đối tượng, nhưng chúng có phạm vi và ứng dụng khác nhau dựa trên tính chất và cấu trúc của dữ liệu.

2.2. Thực hiện khai thác dữ liệu

2.2.1. Sử dụng các phương pháp khai phá dữ liệu đã biết để khai thác dữ liệu đã chọn trong phần 1 (tập phổ biến, phân lớp, phân cụm) với yêu cầu thực hiện tối thiểu 2 phương pháp bất kỳ do SV tự chọn (ví dụ sử dụng Apriori và FP-growth)

2.2.1.1. Thực hiện phân loại bằng Naive Bayes

Ở đây em có thực hiện việc xử lý dữ liệu cho 3 thuộc tính là age, bmi và charges. Chuyển đổi các giá trị về các khoảng để có thể dễ dàng thực hiện được các phần như phân loại và phân cụm ở các phần ở dưới, cụ thể việc thay thế các giá trị thành các khoảng như sau:

Thuộc tính age

Giá trị	Thay thế bởi
18-34	Youth
35-49	Middle-aged
50-64	Senior

Bảng 20 Thay thế giá trị thuộc tính age

Bmi

Giá trị	Thay thế bởi
$15.96 \leq BMI < 18.5$	Underweight
$18.5 \leq BMI < 24.9$	Normal
$25 \leq BMI < 29.9$	Overweight
$30 \leq BMI < 34.9$	Obesisty I
$35 \leq BMI < 39.9$	Obesisty II
$BMI \geq 40$	Obesisty III

Bảng 21 Thay thế giá trị thuộc tính bmi

Charges

Giá trị	Thay thế bởi
0 - 5000	Very Low
5000 - 10000	Low
10000 - 20000	Medium

20000 - 40000	High
Trên 40000	Very High

Bảng 22 Thay thế giá trị thuộc tính charges

Ở đây, em sẽ thực hiện phương pháp phân loại bằng Naive Bayes để dự đoán nhãn cho thuộc tính Smoker.

Smoker	Số lượng
Yes	274
No	1064

Bảng 23 Số lượng giá trị yes và no của thuộc tính smoker

Thực hiện Naive Bayes bằng công thức:

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

Hình 2.17 Công thức tính Naive bayes

2.2.1.1.1. Với giá trị input không trùng với bất kỳ dòng nào trong dataset là:

X = (age = Senior , sex = male , bmi = Obesity II, children = 4 , region = southeast, charges = Medium)

2.2.1.1.2. Xác suất của thuộc tính smoker như sau:

$$P(\text{yes} \mid \text{smoker}) = \frac{274}{1338} = 0{,}2048$$

$$P(\text{no} \mid \text{smoker}) = \frac{1064}{1338} = 0{,}7952$$

2.2.1.1.3. Thực hiện tính các giá trị xác suất có điều kiện:

Age

(Age = Senior có 385 dòng)

$$P(\text{age} = \text{Senior} \mid \text{smoker} = \text{yes}) = \frac{68}{274} = 0{,}2482$$

$$P(\text{age} = \text{Senior} \mid \text{smoker} = \text{no}) = \frac{317}{1064} = 0{,}2979$$

Sex

(Sex = male có 676 dòng)

P(sex = male | smoker = yes) = $\frac{159}{274}$ = 0,5803

P(sex = male | smoker = no) = $\frac{517}{1064}$ = 0,4859

Bmi

(Bmi = Obesity II có 226 dòng)

P(bmi = Obesity II | smoker = yes) = $\frac{52}{274}$ = 0,1898

P(bmi = Obesity II | smoker = no) = $\frac{174}{1064}$ = 0,1635

Children

(Children = 4 có 25 dòng)

P(children = 4 | smoker = yes) = $\frac{3}{274}$ = 0,0109

P(children = 4 | smoker = no) = $\frac{22}{1064}$ = 0,0207

Region

(Region = southeast có 364 dòng)

P(region = southeast | smoker = yes) = $\frac{91}{274}$ = 0,3321

P(region = southeast | smoker = no) = $\frac{273}{1064}$ = 0,2566

Charges

(Charges = Medium có 353 dòng)

P(charges = Medium | smoker = yes) = $\frac{62}{274}$ = 0,2263

P(charges = Medium | smoker = no) = $\frac{291}{1064}$ = 0,2735

P(Smoker = yes | X) = [P(age = Senior | smoker = yes) * P(sex = male | smoker = yes) * P(bmi = Obesity II | smoker = yes) * P(children = 4 | smoker = yes) * P(region

= southeast | smoker = yes) * P(charges = Medium | smoker = yes) * P(Smoker = yes)] **/ P(X)**

$$= \frac{0{,}2482 * 0{,}5803 * 0{,}1898 * 0{,}0109 * 0{,}3321 * 0{,}2263 * 0{,}2048}{P(X)} = 0{,}000004586$$

P(Smoker = no | X) = [P(age = Senior | smoker = no) * P(sex = male | smoker = no) * P(bmi = Obesity II | smoker = no) * P(children = 4 | smoker = no) * P(region = southeast | smoker = no) * P(charges = Medium | smoker = no) * P(Smoker = no)] / **P(X)**

$$= \frac{0{,}2979 * 0{,}4859 * 0{,}1635 * 0{,}0207 * 0{,}2566 * 0{,}2735 * 0{,}7952}{P(X)} = 0{,}00002734$$

2.2.1.1.4. Kết luận:

Vậy ta có thể suy ra với một dòng dữ liệu mới được đưa vào là X thì ta sẽ có nhãn của thuộc tính smoker của dòng này là **no** (vì P(Smoker = yes | X) < P(Smoker = no | X))

2.2.1.2. Thực hiện phân cụm với K-Means

Để thực hiện phân cụm bằng K-means thủ công ở Excel, em chuyển đổi các giá trị của thuộc tính danh nghĩa về số, sau đó thực hiện các bước như sau:

Bước 1: Đầu tiên em sẽ chọn giá trị **k** để có thể thực hiện được việc chọn số centroid của dataset. Và giá trị **k** em chọn là bằng 3 vì vậy em sẽ chọn ngẫu nhiên 3 dòng centroid từ dataset ra làm centroid.

Bước 2: Sau khi đã lấy ra được 3 centroid, em thực hiện tính khoảng cách của từng dòng (từng datapoint) trong dữ liệu đến từng centroid **bằng khoảng cách Manhattan.**

Bước 3: Tiếp đến, sau khi đã tính khoảng cách của từng dòng đến 3 centroid, em thực hiện việc phân cụm ra bằng cách xét giá trị nào nhỏ nhất sẽ nằm ở cụm đã được phân. Ví dụ, em có bảng gồm các cột và 2 dòng với các giá trị như sau:

Cụm / Dòng	1	2	3	Thuộc
Giá trị datapoint 1	3	5	7	Cụm 1 (vì 3 là giá trị nhỏ nhất và nằm ở cụm 1)
Giá trị datapoint 2	4	8	1	Cụm 3 (vì 1 là giá trị nhỏ nhất và nằm ở cụm 3)

Bảng 24 Ví dụ bước 3 phân cụm

Bước 4: Sau khi đã phân cụm cho các datapoint ở lần đầu tiên, em thực hiện tính lại 3 centroid bằng cách dùng hàm **AVERAGE()** trong excel để tính lại 3 centroid cho

dataset, các lần phân cụm tiếp theo cũng cần tính lại centroid sau khi đã phân cụm trước đó.

Bước 5: Sau khi đã thực hiện phân cụm cho các datapoint, em thực hiện việc so sánh cụm đã phân so với lần trước đó (bỏ qua lần đầu tiên). Nếu cụm của các datapoint được phân ở lần mới nhất trùng với cụm của các datapoint ở lần trước đó thì em sẽ dừng việc phân cụm. Còn nếu các cụm vẫn còn khác nhau thì em vẫn tiếp tục phân cụm cho những lần tiếp theo và sau đó so sánh các cụm.

Ở dataset của em, em dừng phân cụm ở lần thứ 15 vì chỉ còn 1 dòng khác cụm so với lần thứ 14.

Để xem cụ thể phần phân cụm trong excel, em có đính kèm link excel ở phần phụ lục, thầy có thể xem qua ạ.

2.2.2. Thực hiện đánh giá các mẫu thu được bằng các phương pháp đã biết bằng cách chọn 2 trong số các phương pháp đánh giá để đánh giá kết quả của việc thực hiện ở phần 2.2.1

2.2.2.1. Đánh giá phân loại Naive Bayes

Để thực hiện đánh giá phân loại bằng Naive Bayes, em thực hiện các bước như sau:

Bước 1: Em sẽ chia tập train và tập test theo tỷ lệ 8:2 để dự đoán nhãn của thuộc tính smoker bên tập test. Như vậy, tập train sẽ là 1070 dòng đầu của dataset và tập test sẽ là 268 dòng cuối cùng của dataseet.

Bước 2: Sau đó em sẽ dùng tập train để tính các giá trị xác suất của từng giá trị của mỗi thuộc tính. Ví dụ em có 1 cột như sau ở tập train:

Cột	Smoker
A	Yes
B	No
C	Yes
A	Yes
B	No

Bảng 25 Ví dụ bước 2 đánh giá phân loại

Em thực hiện tính P(A | smoker = yes) và P(a | smoker = no). Tương tự cho các giá trị khác trong cột và các cột thuộc tính khác trong dataset. Sau đó, lưu các giá trị P(X) để thực hiện dự đoán nhãn bên tập test.

Bước 3: Sau khi đã có các giá trị bên tập train, em tiến hành dự đoán nhãn cho từng dòng bên tập test. Sau khi đã dự đoán nhãn cho các dòng xong, em thực hiện việc so sánh các nhãn vừa được dự đoán so với nhãn đã có trước đó ở tập train xem có trùng khớp hay không, nếu trùng em sẽ cho nó là 1 còn không trùng sẽ là 0.

Bước 4: Tiếp đó, em đếm số lượng 4 giá trị là TP, FN, FP, TN để tiến hành hoàn thành Confusion matrix cho việc đánh giá.

Để xem cụ thể phần đánh giá phân loại bằng Naive Bayes trong excel, em cũng có đính kèm link excel ở phần phụ lục, thầy có thể xem qua ạ.

2.2.2.2. Đánh giá phân cụm K-means

Phần thực hiện đánh giá phân cụm này do không đủ thời gian nên em làm không kịp, mong thầy thông cảm ạ.

CHƯƠNG 3: PHÂN TÍCH – THỐNG KÊ BẰNG PYTHON TRÊN CƠ SỞ DỮ LIỆU ĐÃ CHỌN

3.1. Thực hiện lại phần 1 và phần 2 với Python

3.1.1. Phần 1

3.1.1.1. Đọc file và hiển thị ra tổng số dòng và cột của dataset

```
Tổng số dòng của dataset: 1338
Tổng số cột của dataset: 7
```

Hình 3.1 Đọc file và hiển thị số cột và số dòng

3.1.1.2. Kiểm tra số giá trị null (missing data)

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Hình 3.2 Số giá trị null

3.1.1.3. Kiểu dữ liệu của các thuộc tính

```
Kiểu dữ liệu của mỗi cột:
 age           int64
sex           object
bmi          float64
children       int64
smoker        object
region        object
charges      float64
dtype: object
```

Hình 3.3 Kiểu dữ liệu của các thuộc tính

3.1.1.4. Kiểm tra các giá trị nhị phân hoặc rời rạc kèm theo số lượng giá trị của từng thuộc tính

3.1.1.4.1. Thuộc tính sex

```
Cột sex:
Giá trị:
 sex
male      676
female    662
Name: count, dtype: int64
Tỷ lệ phần trăm:
 sex
male      50.523169
female    49.476831
Name: proportion, dtype: float64
Mode:
 male
```

Hình 3.4 Chi tiết thuộc tính sex

3.1.1.4.2. Thuộc tính children

```
Cột children:
Giá trị:
 children
0    574
1    324
2    240
3    157
4     25
5     18
Name: count, dtype: int64
Tỷ lệ phần trăm:
 children
0    42.899851
1    24.215247
2    17.937220
3    11.733931
4     1.868460
5     1.345291
Name: proportion, dtype: float64
Mode:
 0
```

Hình 3.5 Chi tiết thuộc tính children

3.1.1.4.3. Thuộc tính smoker

```
Cột smoker:
Giá trị:
 smoker
no     1064
yes     274
Name: count, dtype: int64
Tỷ lệ phần trăm:
 smoker
no     79.521674
yes    20.478326
Name: proportion, dtype: float64
Mode:
 no
```

Hình 3.6 Chi tiết thuộc tính smoker

3.1.1.4.4. Thuộc tính region

```
Cột region:
Giá trị:
 region
southeast    364
southwest    325
northwest    325
northeast    324
Name: count, dtype: int64
Tỷ lệ phần trăm:
 region
southeast    27.204783
southwest    24.289985
northwest    24.289985
northeast    24.215247
Name: proportion, dtype: float64
Mode:
 southeast
```

Hình 3.7 Chi tiết thuộc tính region

3.1.1.5. Kiểm tra các giá trị min, max, median, mean,..của các thuộc tính số

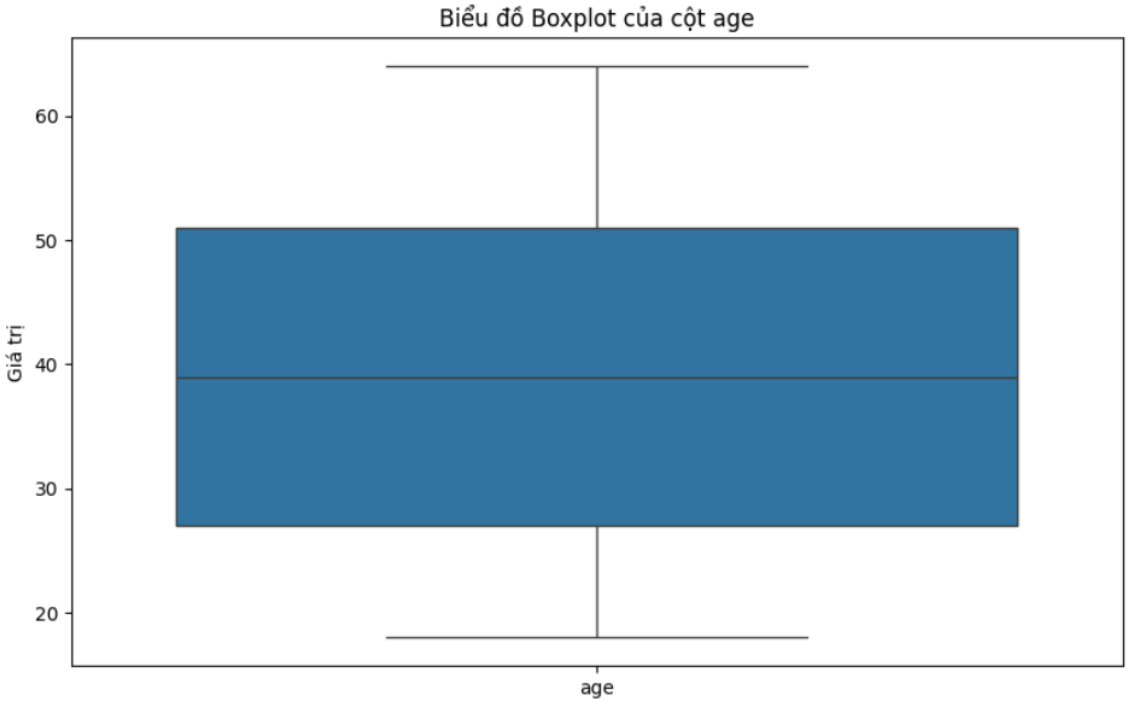
Các giá trị thống kê cho các thuộc tính kiểu số:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010
median	39.000000	30.400000	1.000000	9382.033000
midrange	41.000000	34.545000	2.500000	32446.150955
mode	18.000000	32.300000	0.000000	1639.563100

Hình 3.8 Chi tiết các giá trị của từng thuộc tính số

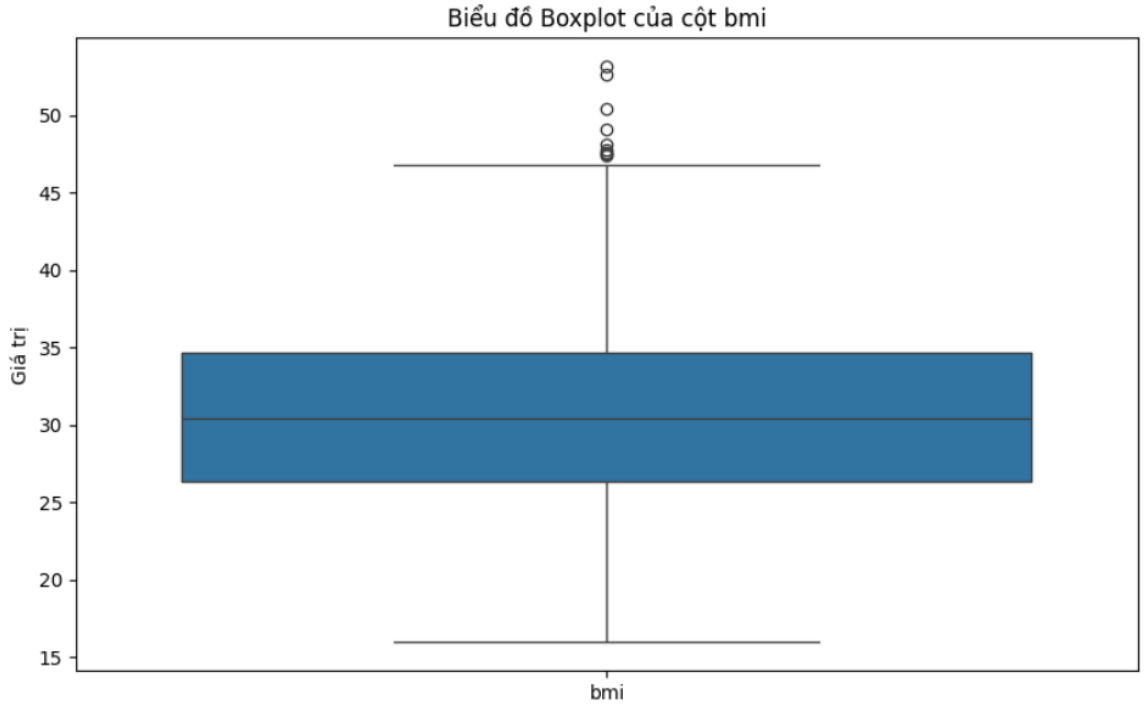
3.1.2. Phần 2

3.1.2.1. Boxplot của cột age



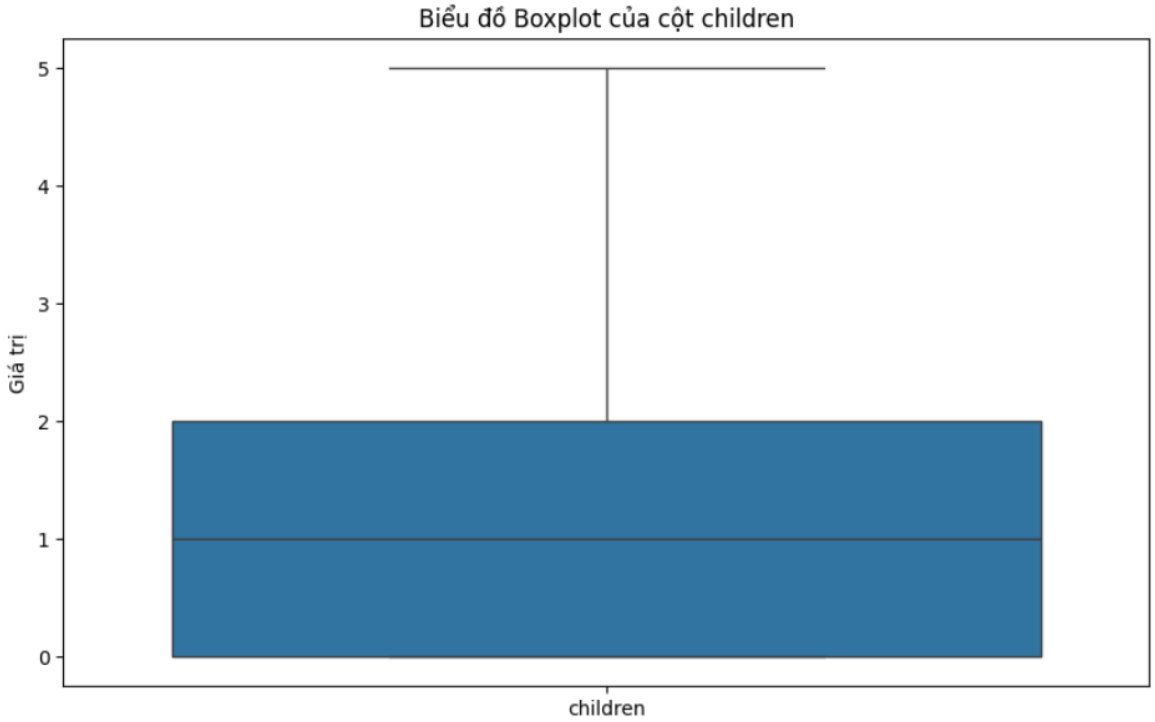
Hình 3.9 Boxplot của cột age

3.1.2.2. Boxplot của cột bmi



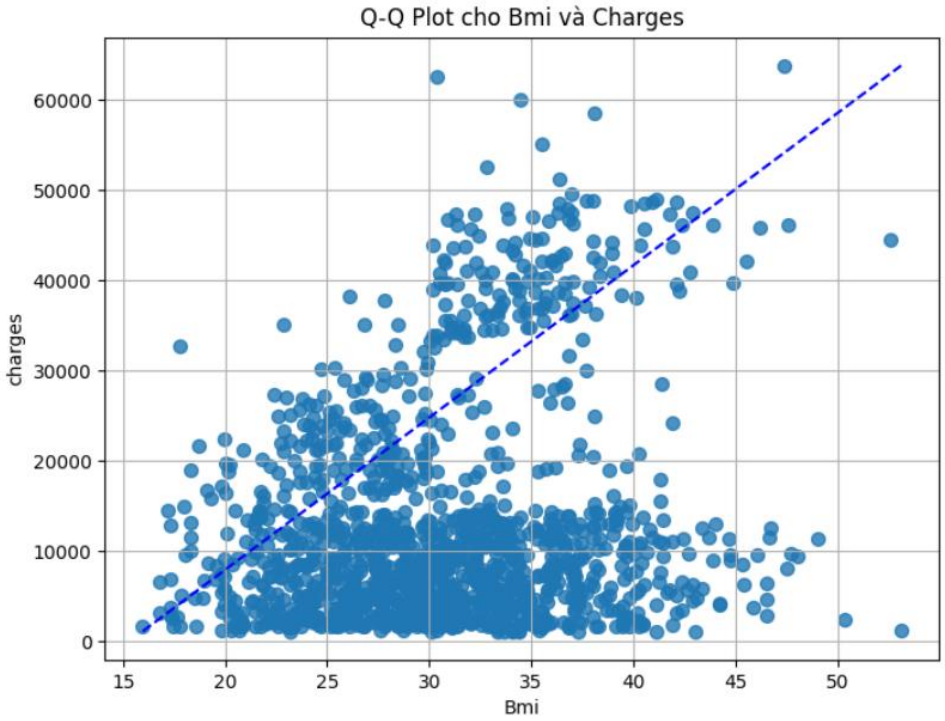
Hình 3.10 Boxplot của cột bmi

3.1.2.3. Boxplot của cột children



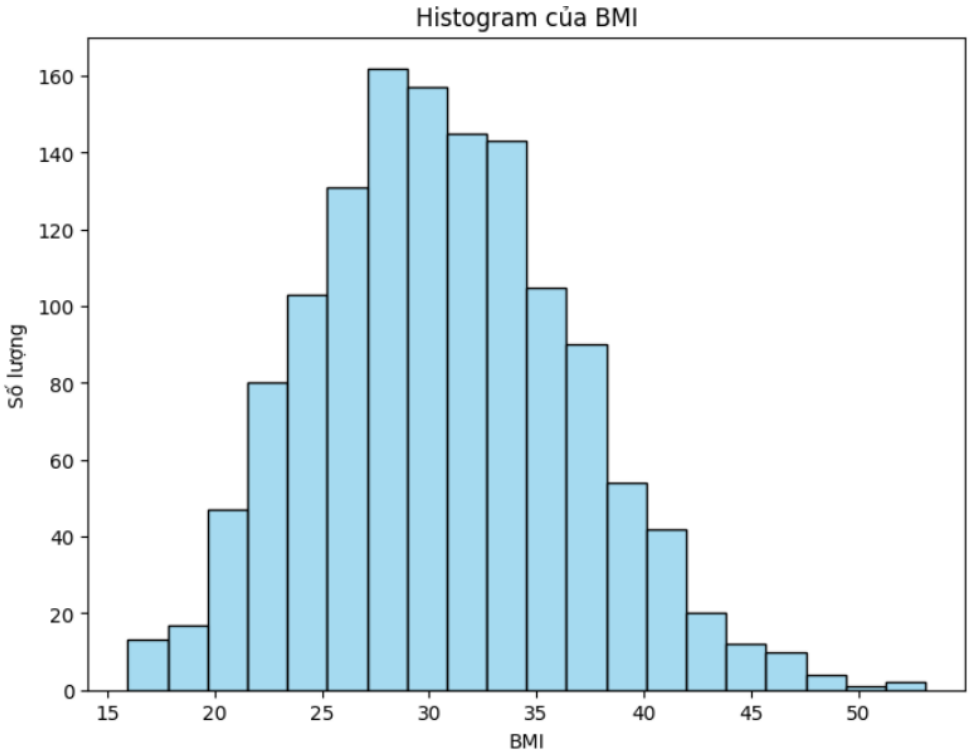
Hình 3.11 Boxplot của cột children

3.1.2.4. Biểu đồ Q-Q Plot của cột bmi và charges



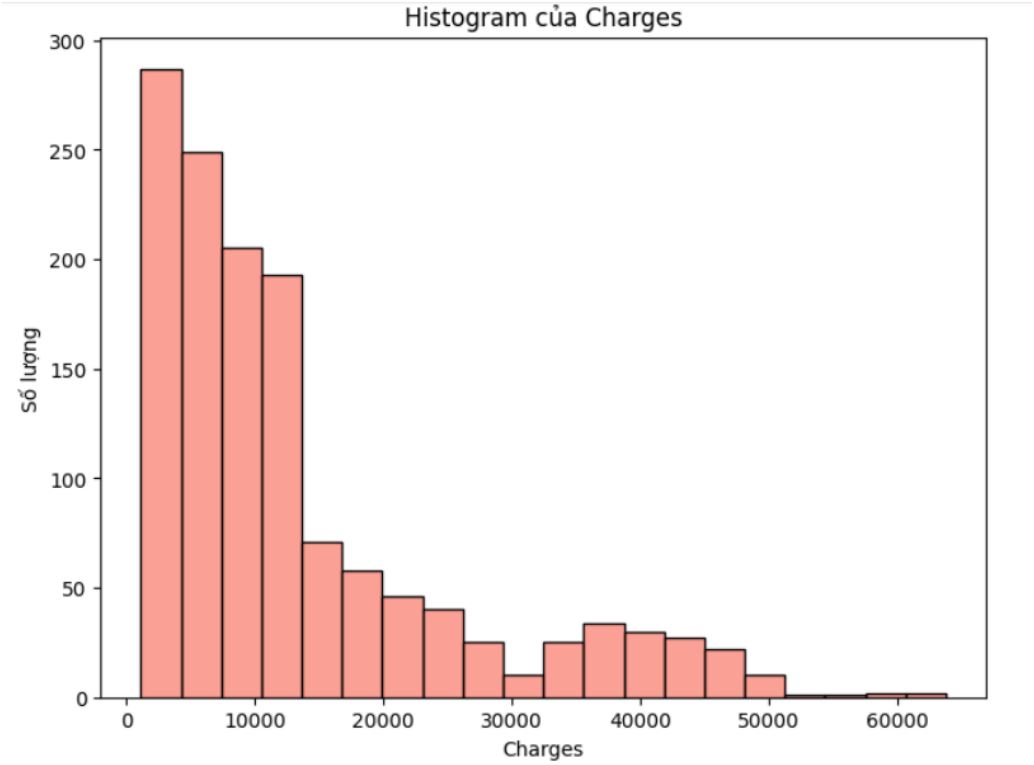
Hình 3.12 Q-Q Plot của cột bmi và charges

3.1.2.5. Biểu đồ Histogram của cột bmi



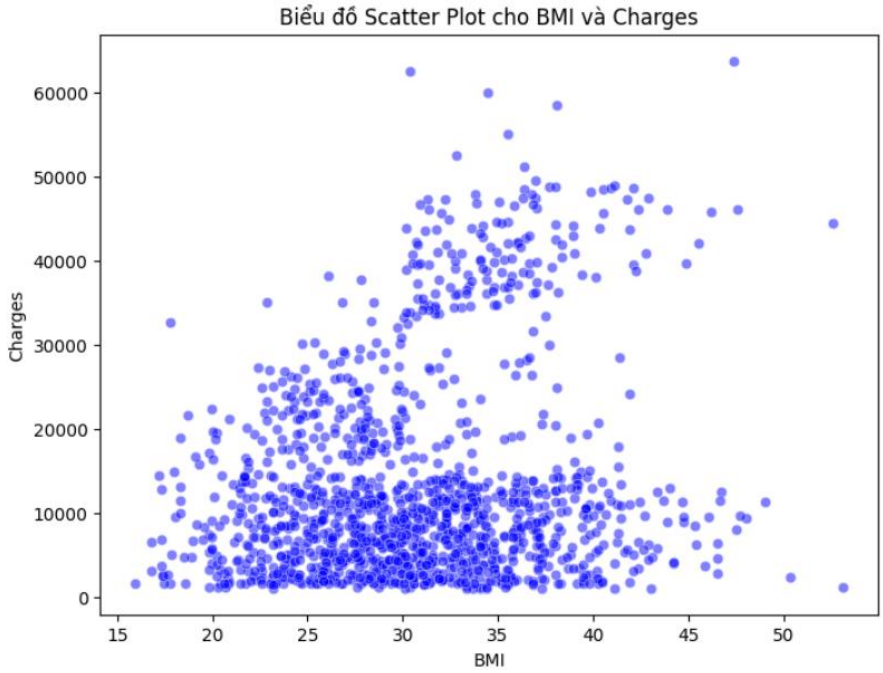
Hình 3.13 Histogram của bmi

3.1.2.6. Biểu đồ Histogram của cột charges



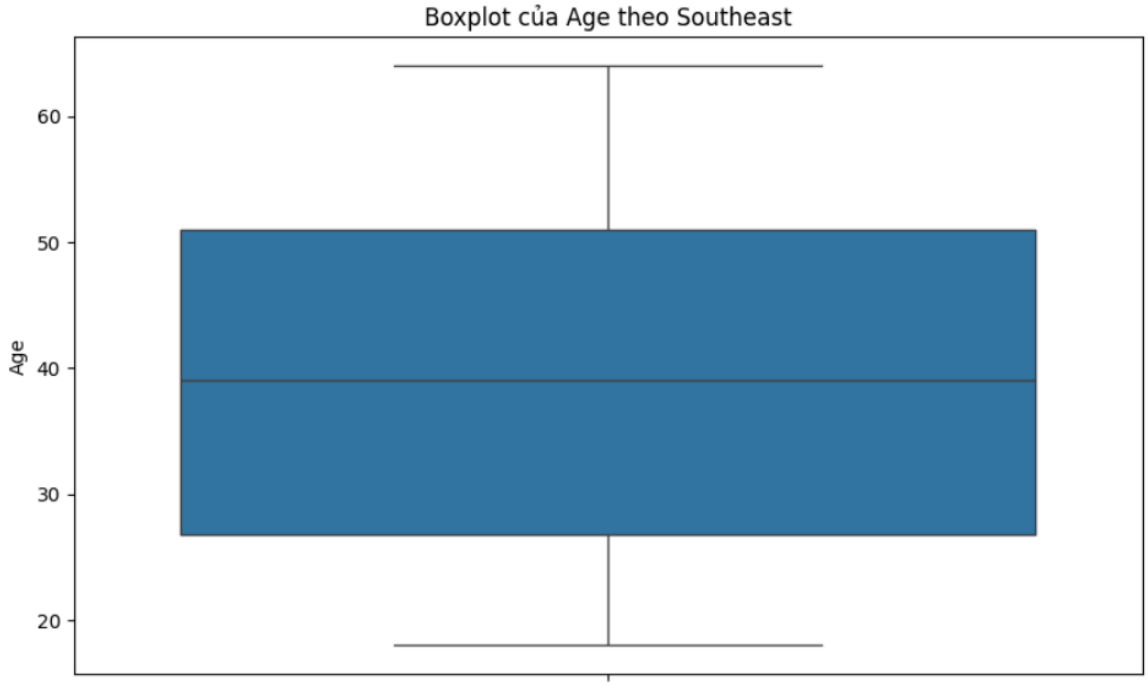
Hình 3.14 Histogram của charges

3.1.2.7. Biểu đồ phân tán của cột bmi và cột charges



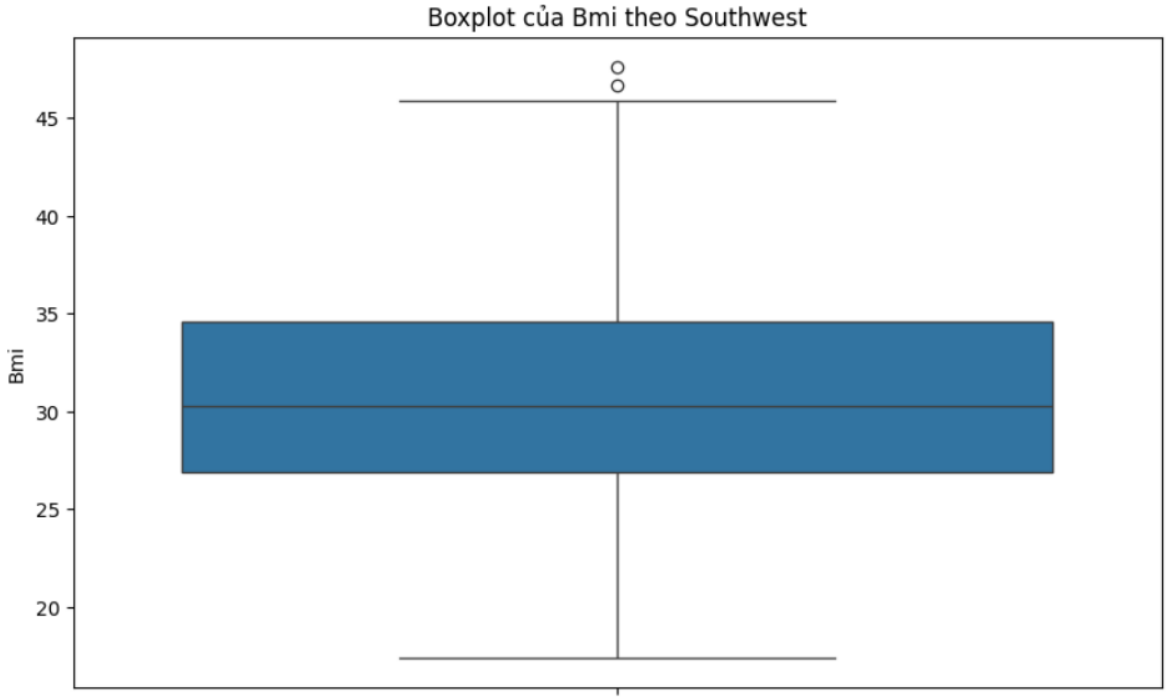
Hình 3.15 Scatter plot của bmi và charges

3.1.2.8. Boxplot của age lọc theo danh nghĩa southeast



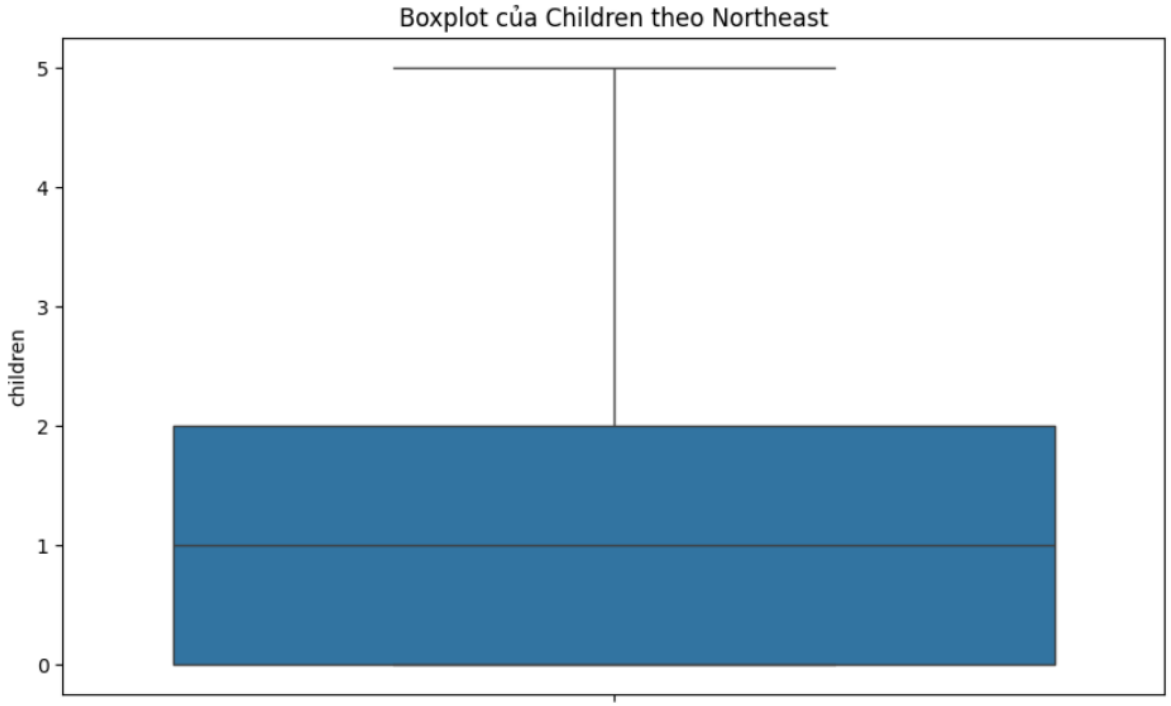
Hình 3.16 Boxplot của age theo southeast

3.1.2.9. Boxplot của bmi lọc theo danh nghĩa southwest



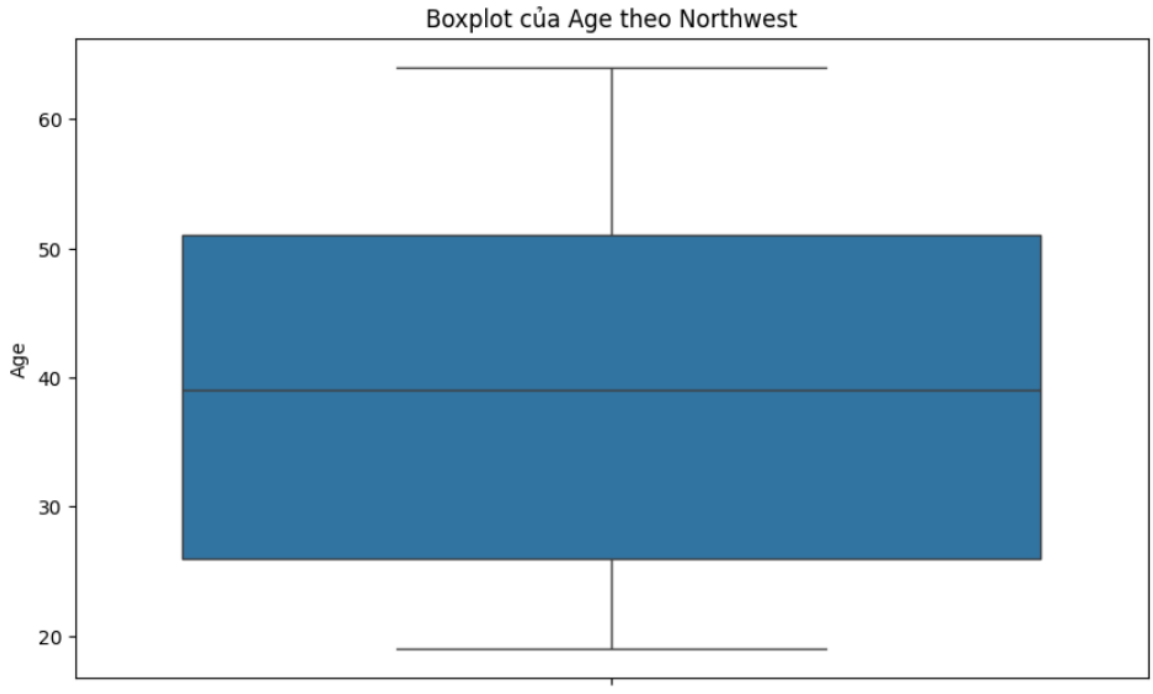
Hình 3.17 Boxplot của bmi theo southwest

3.1.2.10. Boxplot của children lọc theo danh nghĩa northeast



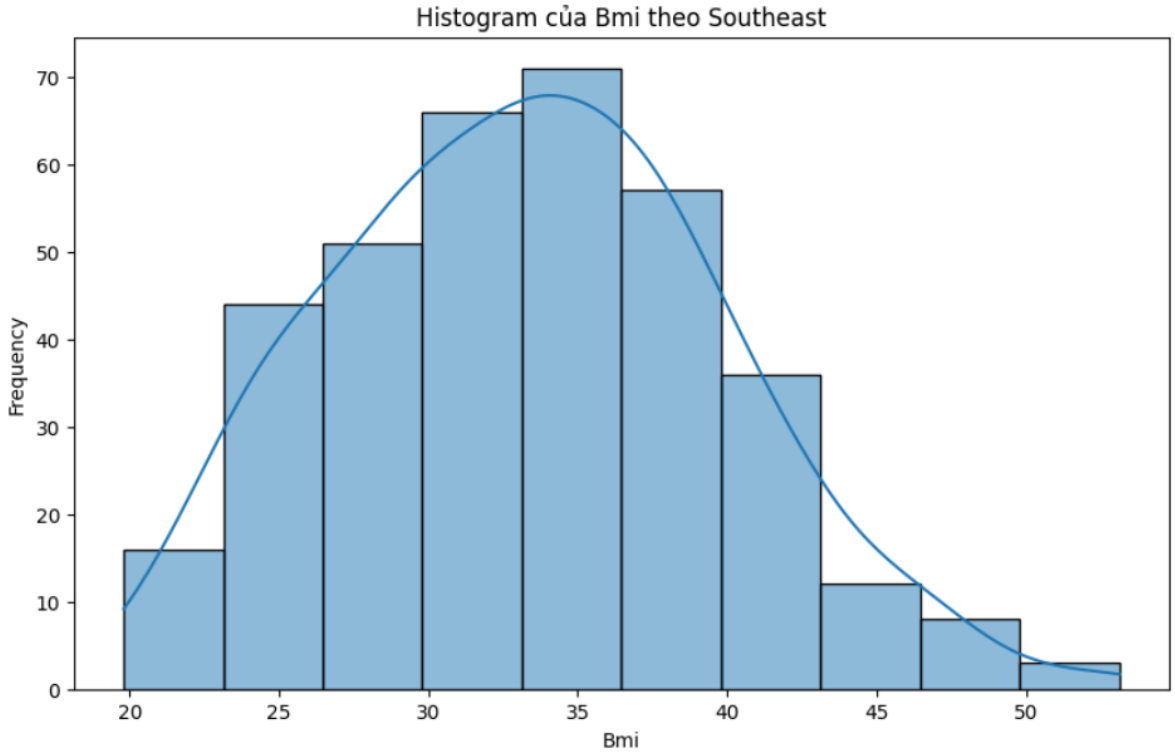
Hình 3.18 Boxplot của children theo northeast

3.1.2.11. Boxplot của age lọc theo danh nghĩa northwest



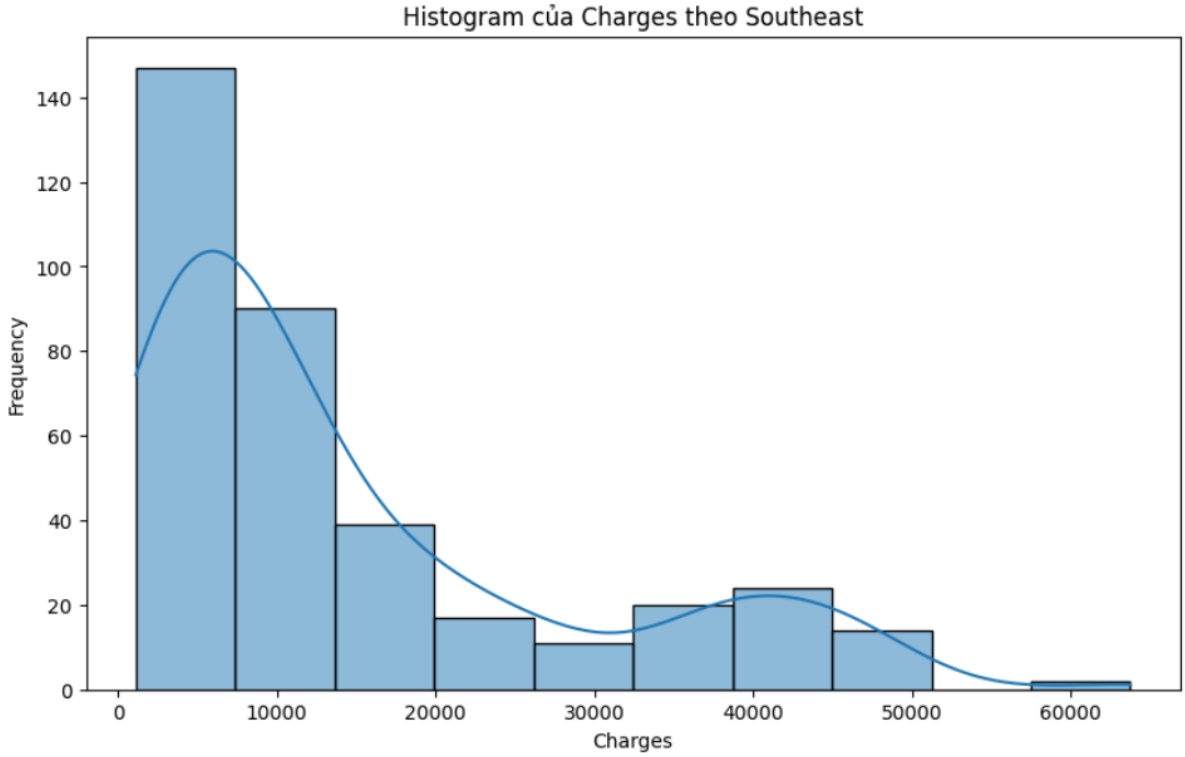
Hình 3.19 Boxplot của age theo northwest

3.1.2.12. Histogram của bmi lọc theo danh nghĩa southeast



Hình 3.20 Histogram của bmi theo southeast

3.1.2.13. Histogram của charges lọc theo danh nghĩa southeast

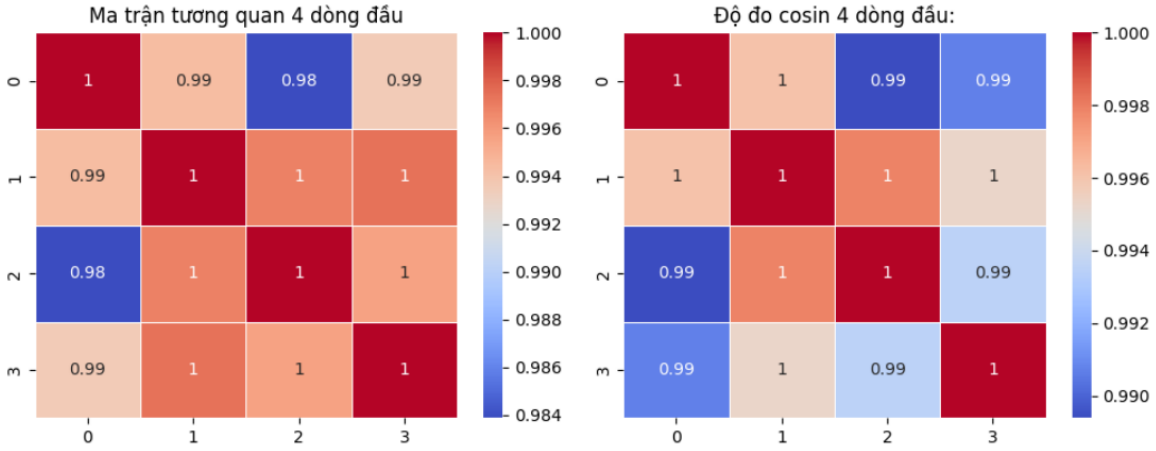


Hình 3.21 Histogram của charges theo southeast

3.1.2.14. Ma trận tương quan và độ đo cosin của 4 dòng dữ liệu đầu tiên

```
Ma trận tương quan:
          0         1         2         3
0  1.000000  0.994393  0.983882  0.993578
1  0.994393  1.000000  0.996885  0.997393
2  0.983882  0.996885  1.000000  0.995643
3  0.993578  0.997393  0.995643  1.000000

Độ đo cosin:
          0         1         2         3
0  1.000000  0.996080  0.989387  0.990696
1  0.996080  1.000000  0.997916  0.995257
2  0.989387  0.997916  1.000000  0.993502
3  0.990696  0.995257  0.993502  1.000000
```



Hình 3.22 Ma trận tương quan và độ đo cosin

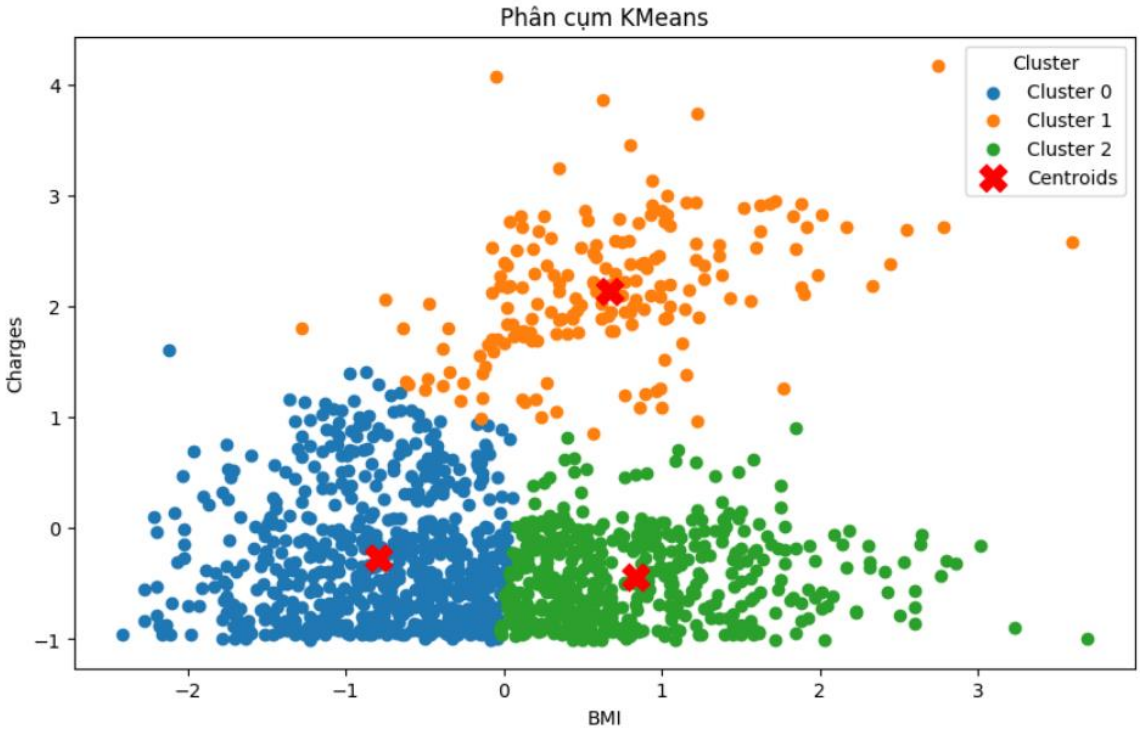
3.1.2.15. Phân loại bằng Naive Bayes cho nhãn smoker

```
# In kết quả dự đoán
print('Nhãn smoker được dự đoán cho input X = (age = Senior , sex = male , bmi = Obesity II, children = 4 , region = southeast, charges = Medium) là: ',predicted_label
```

```
Nhãn smoker được dự đoán cho input X = (age = Senior , sex = male , bmi = Obesity II, children = 4 , region = southeast, charges = Medium) là:  no
```

Hình 3.23 Kết quả phân loại bằng Naive Bayes

3.1.2.16. Phân cụm với K-means cho 2 thuộc tính bmi và charges



Hình 3.24 Kết quả phân cụm

3.1.2.17. Đánh giá kết quả phân loại bằng Confusion matrix

```
Confusion Matrix:
[[209   0]
 [ 59   0]]

Confusion Matrix Values:
{'True Positives (TP)': 0, 'False Negatives (FN)': 59, 'False Positives (FP)': 0, 'True Negatives (TN)': 209}

Classification Report:
              precision    recall  f1-score   support

          no       0.78      1.00      0.88       209
         yes       0.00      0.00      0.00        59

    accuracy                           0.78       268
   macro avg       0.39      0.50      0.44       268
weighted avg       0.61      0.78      0.68       268
```

Hình 3.25 Kết quả đánh giá phân loại

3.2. So sánh kết quả thực hiện của việc sử dụng công cụ với việc thủ công (phần 2.1 và 2.2)

3.2.1. Về độ chính xác và kết quả

Kết quả của hai phương pháp không khác nhau quá nhiều, chứng tỏ cả hai phương pháp đều có độ chính xác gần tương đồng nhau khi xử lý dữ liệu. Kết quả và độ chính xác của việc xử lí bằng Python sẽ cao hơn so với việc xử lí ở Excel. Cụ thể:

Về việc vẽ các biểu đồ: Python và Excel đều cho ra các biểu đồ tương tự nhau, độ chính xác khá cao.

Về việc tính toán ma trận tương quan và độ đo cosin: Python cho ra kết quả có chênh lệch một chút so với việc xử lí bằng Excel nhưng không đáng kể.

Về việc phân loại bằng Naive Bayes: cả 2 cách đều cho ra kết quả giống nhau.

Về việc phân cụm: ở excel em thực hiện phân cụm được trên tất cả các cột, còn ở Python chỉ thực hiện được với 2 thuộc tính cụ thể, nên không thể so sánh kết quả chính xác được.

3.2.2. Về tốc độ và hiệu suất

Về thời gian, chắc chắn việc xử lí bằng Python sẽ nhanh hơn rất nhiều so với việc làm bằng tay ở Excel, đặc biệt khi làm việc với tập dữ liệu lớn. Hiệu suất của việc xử lí bằng Python cũng tốt hơn nhiều so với Excel.

3.2.3. Kết luận

Cả hai phương pháp đều có ưu và nhược điểm riêng. Excel phù hợp với các tác vụ đơn giản, dễ thực hiện với các tập dữ liệu nhỏ và không yêu cầu nhiều kiến thức lập trình. Python vượt trội hơn trong việc xử lý dữ liệu lớn, tự động hóa các quy trình và thực hiện các phân tích phức tạp.

Đối với các nhiệm vụ yêu cầu phân tích nhanh chóng và đơn giản, Excel là lựa chọn hợp lý. Tuy nhiên, với các dự án yêu cầu xử lý dữ liệu lớn, phức tạp và cần tự động hóa, Python là lựa chọn tối ưu hơn.

CHƯƠNG 4: PHỤ LỤC

4.1. File excel phân cụm

Clustering.xlsx

4.2. File excel đánh giá phân loại

Evaluate.xlsx

TÀI LIỆU THAM KHẢO

[1]. *Tài liệu môn học Khoa Học Dữ Liệu*, thầy Lê Văn Hạnh.

[2]. Medical Cost Personal Datasets (kaggle.com)

[3]. Orange Data Mining