Đại Học Quốc Gia Thành phố Hồ Chí Minh
Trường Đại Học Công Nghệ Thông Tin

Phân tích ảnh hưởng của các chỉ số sức khỏe đến tiến triển bệnh đái tháo đường

DS304.L21 – Phân tích và thiết kế thực nghiệm

NHÓM 9

Giảng viên hướng dẫn

TS. Đỗ Trọng Hợp

Nhóm sinh viên thực hiện

Thái Minh Triết **19522397**
Chu Hà Thảo Ngân **19521882**
Võ Tuấn Anh **19521226**

Nội dung thuyết trình

1. Giới thiệu chung
2. Tổng quan bộ dữ liệu
3. Tiền xử lý dữ liệu
4. Phân tích hồi quy
5. Đánh giá mô hình
6. Kết luận

Giới thiệu chung

Giới thiệu chung

1. Đặt vấn đề

- Đái tháo đường được xem là một căn bệnh phổ biến trong cộng đồng.
- Ở Mỹ có xu hướng gia tăng ngày càng cao với con số ước tính lên đến 79 triệu người trưởng thành mắc bệnh.
- Trong đó 50% người mắc đái tháo đường không biết mình có bệnh vì chưa có triệu chứng rõ ràng.

Giới thiệu chung

1. Đặt vấn đề

- Đái tháo đường là một căn bệnh rối loạn chuyển hóa trong hormone tuyến tuỵ dẫn đến thiếu hụt lượng insuline trong máu.
- Các tiêu chuẩn để chẩn đoán bệnh đái tháo đường hiện nay cần dựa vào nhiều chỉ số và quá trình phức tạp. Vì thế việc chẩn đoán phụ thuộc vào các chỉ số này từ cơ thể.
- Cần phải có bộ dữ liệu với các chỉ số y tế liên quan để đánh giá phần nào tình hình của bệnh, giúp tăng hiệu quả và giảm đi chi phí cho quá trình chuẩn đoán.

Giới thiệu chung

2. Mục tiêu nghiên cứu

- Phân tích ảnh hưởng giữa các yếu tố là chỉ số y tế liên quan lên tiến triển bệnh.
- Xem xét sự tương tác giữa các yếu tố được thu thập từ các bệnh nhân bộ dữ liệu có sẵn.
- Xây dựng các mô hình hồi quy và đánh giá kết quả dự đoán của các mô hình.

Giới thiệu chung

3. Ứng dụng

- Cho phép sử dụng trong các chương trình chuẩn đoán sớm bằng các thông số cơ thể. Giúp giảm đi chi phí trong quá trình chuẩn đoán.
- Dùng trong các hệ khuyến nghị đưa ra chế độ ăn uống hoặc điều trị phù hợp với mỗi cá nhân. Giúp gia tăng hiệu quả cho việc điều trị

Tổng quan bộ dữ liệu

Thông tin chung về bộ dữ liệu

❖ Tên bộ dữ liệu

- Diabetes dataset

❖ Nguồn thu nhập

- https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

❖ Kích thước của bộ dữ liệu

- 442 dòng x 11 cột

❖ Tác giả

- Bradley Efron, Trevor Hastie, Iain Johnstone và Robert Tibshirani,

  Đại học Stanford, Hoa Kỳ

Thông tin chung về bộ dữ liệu

❖ Thông tin thuộc tính:

- **AGE**: độ tuổi của bệnh nhân
- **SEX**: giới tính của bệnh nhân (1 - Nam; 2 – Nữ)
- **BMI**: body mass index (chỉ số khối cơ thể)
- **BP**: average blood pressure (giá trị huyết áp trung bình)
- **S1**: tổng lượng cholesterol trong huyết thanh (tc - total serum cholesterol)
- **S2**: giá trị lipoprotein tỷ trọng thấp (ldl - low-density lipoproteins)
- **S3**: giá trị lipoprotein tỷ trọng cao (hdl - high-density lipoproteins)
- **S4**: tỉ lệ giữa cholesterol toàn phần so với lượng HDL (tch - total cholesterol / HDL)
- **S5**: mức triglyceries có trong huyết thanh có thể ghi nhận (ltg - possibly log of serum triglycerides level)
- **S6**: chỉ số mức đường huyết (glu - blood sugar level)
- **Y**: giá trị định lượng về tiến triển bệnh của bệnh nhân sau 1 năm kể từ thời điểm ghi nhận

Trực quan bộ dữ liệu

Thống kê mô tả bộ dữ liệu

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
count	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000	442.000000
mean	48.518100	1.468326	26.375792	94.647014	189.140271	115.439140	49.788462	4.070249	4.641411	91.260181	152.133484
std	13.109028	0.499561	4.418122	13.831283	34.608052	30.413081	12.934202	1.290450	0.522391	11.496335	77.093005
min	19.000000	1.000000	18.000000	62.000000	97.000000	41.600000	22.000000	2.000000	3.258100	58.000000	25.000000
25%	38.250000	1.000000	23.200000	84.000000	164.250000	96.050000	40.250000	3.000000	4.276700	83.250000	87.000000
50%	50.000000	1.000000	25.700000	93.000000	186.000000	113.000000	48.000000	4.000000	4.620050	91.000000	140.500000
75%	59.000000	2.000000	29.275000	105.000000	209.750000	134.500000	57.750000	5.000000	4.997200	98.000000	211.500000
max	79.000000	2.000000	42.200000	133.000000	301.000000	242.400000	99.000000	9.090000	6.107000	124.000000	346.000000

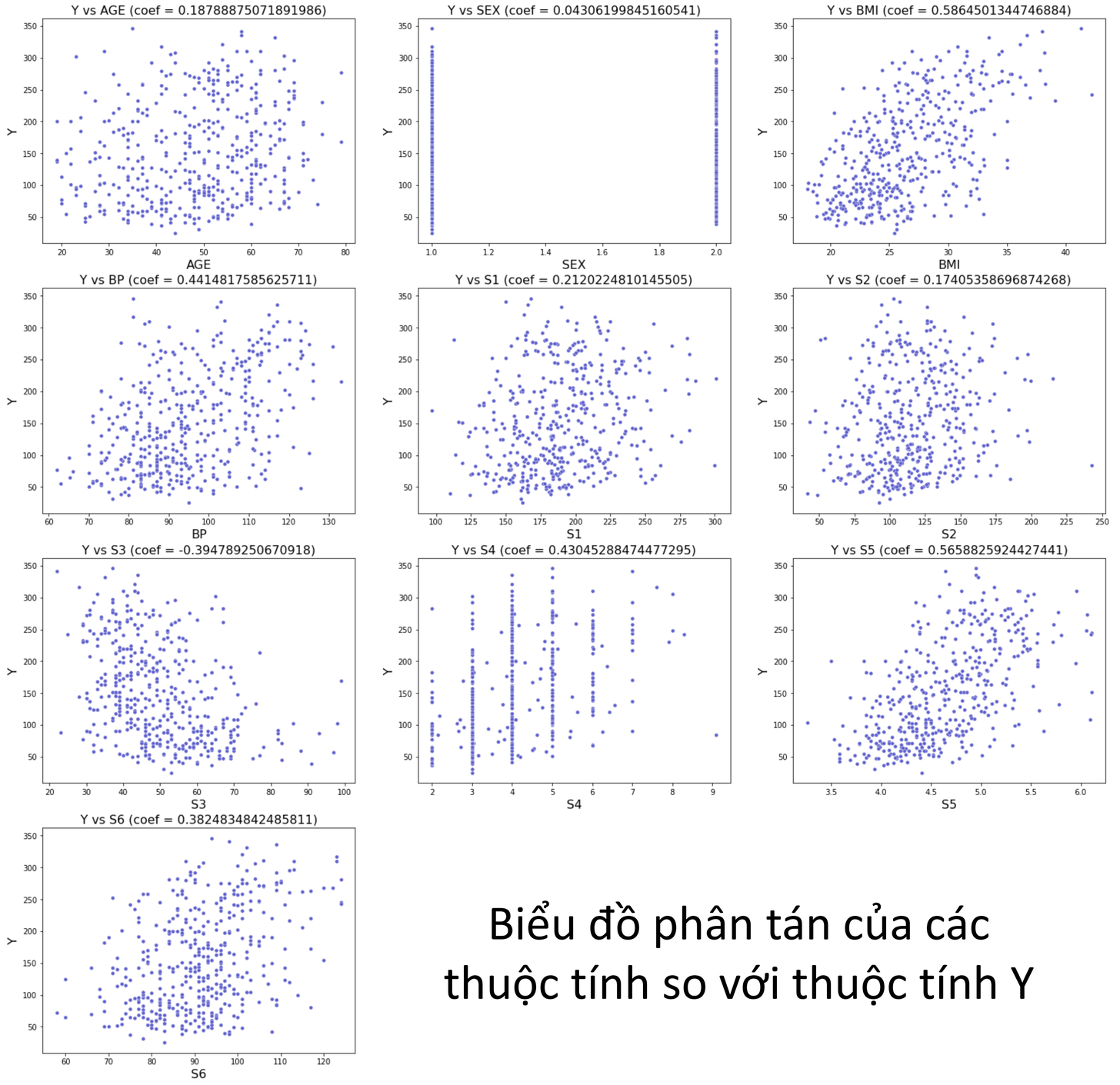
Trực quan bộ dữ liệu

❖ Nhận xét:

- Thuộc tính **BMI, S3, S4 và S5** có phân bố dữ liệu theo xu hướng quan hệ tuyến tính với Y. **BMI, S4 và S5** càng tăng, **S3** càng nhỏ thì khả năng tiến triển bệnh nhanh hơn.

- Các thuộc tính còn lại phân bố hỗn loạn và ít có quan hệ tuyến tính với thuộc tính Y.

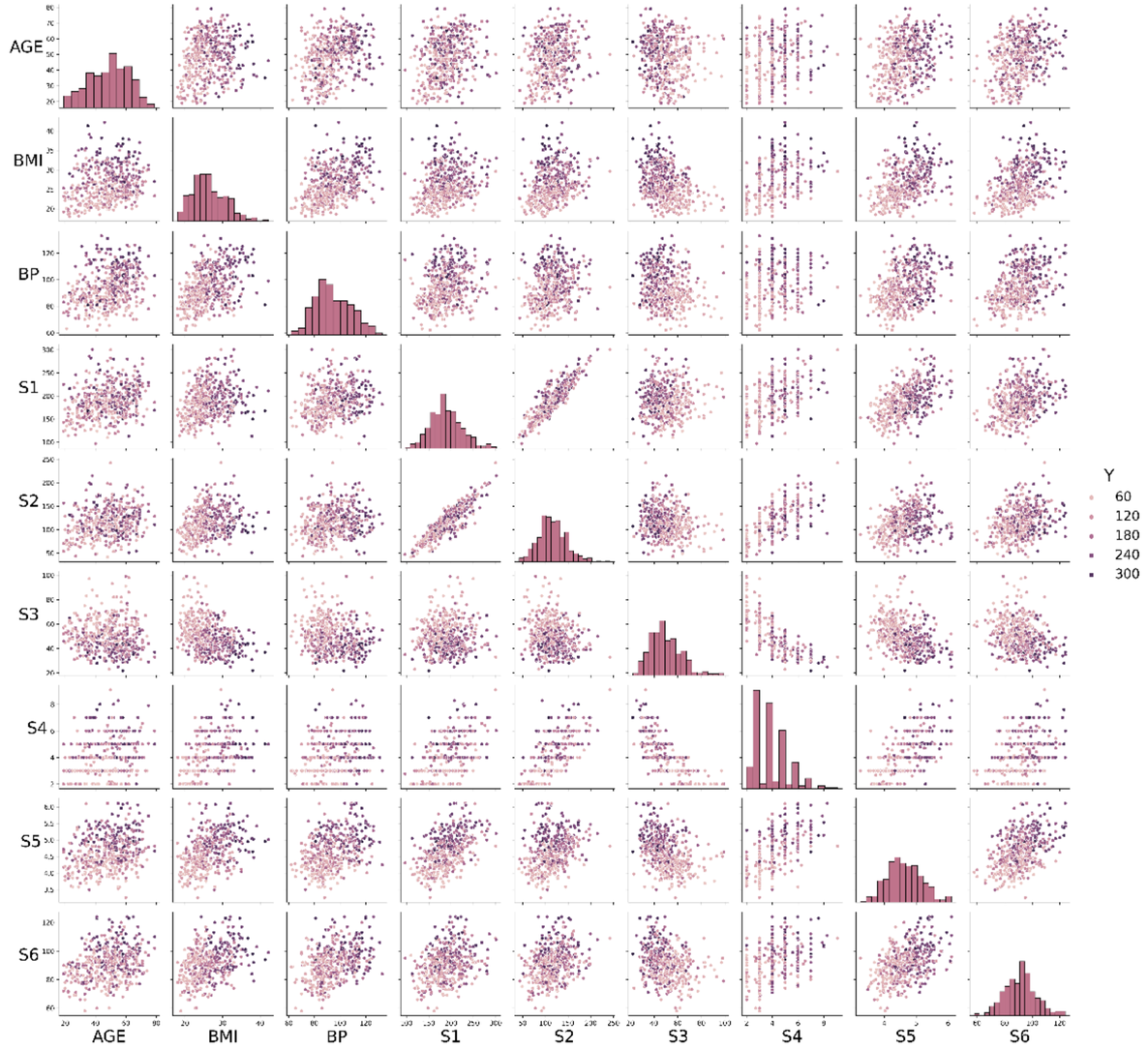
- Thuộc tính **S4 và S2** xuất hiện các điểm dữ liệu ngoại lệ trên biểu đồ, có thể gây ảnh hưởng tới hiệu suất các mô hình hồi quy.



Biểu đồ phân tán của các thuộc tính so với thuộc tính Y

Trực quan bộ dữ liệu

Biểu đồ phân tán giữa các thuộc tính định lượng

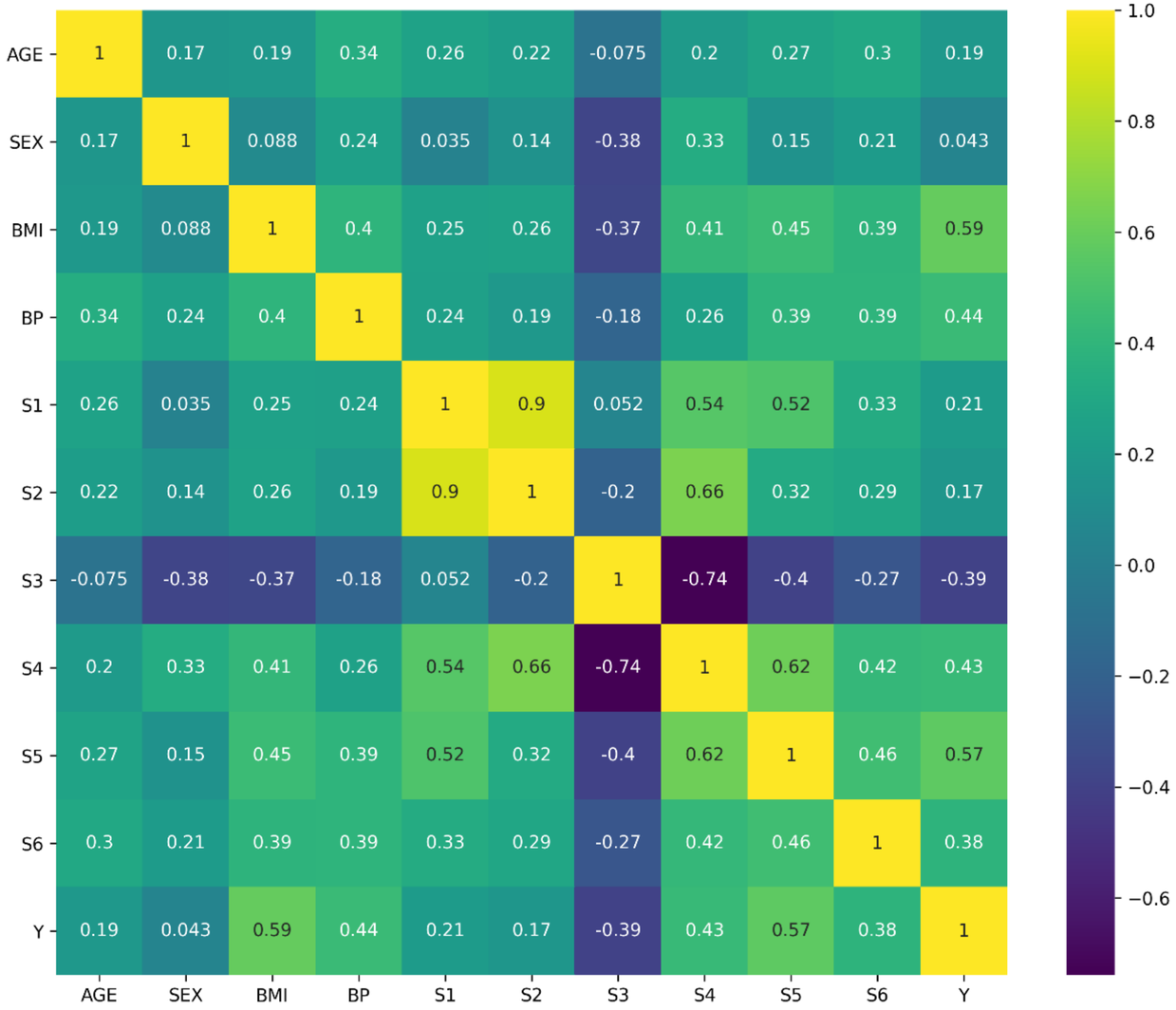


❖ **Nhận xét:**

- Thuộc tính **S2 và S1** thể hiện rõ ràng quan hệ tuyến tính.
- Một số cặp thuộc tính có quan hệ tuyến tính: **S2 và S4**, **S3 và S4**, **S6 và S5**.
- Các biểu đồ còn lại thì phân bố hỗn loạn và không theo một quy luật cụ thể.
- Bệnh nhân có chỉ số BMI và S5 thấp thì tiến triển bệnh chậm, khi hai chỉ số này cùng tăng thì bệnh có khả năng thể tiến triển nhanh hơn. Mặt khác, ở biểu đồ của S3 so với BP, S3 cao nhưng BP thấp thì bệnh tiến triển nhẹ, S3 thấp nhưng BP cao cho thấy bệnh tiến triển nhanh.

Trực quan bộ dữ liệu

Biểu đồ hệ số tương quan

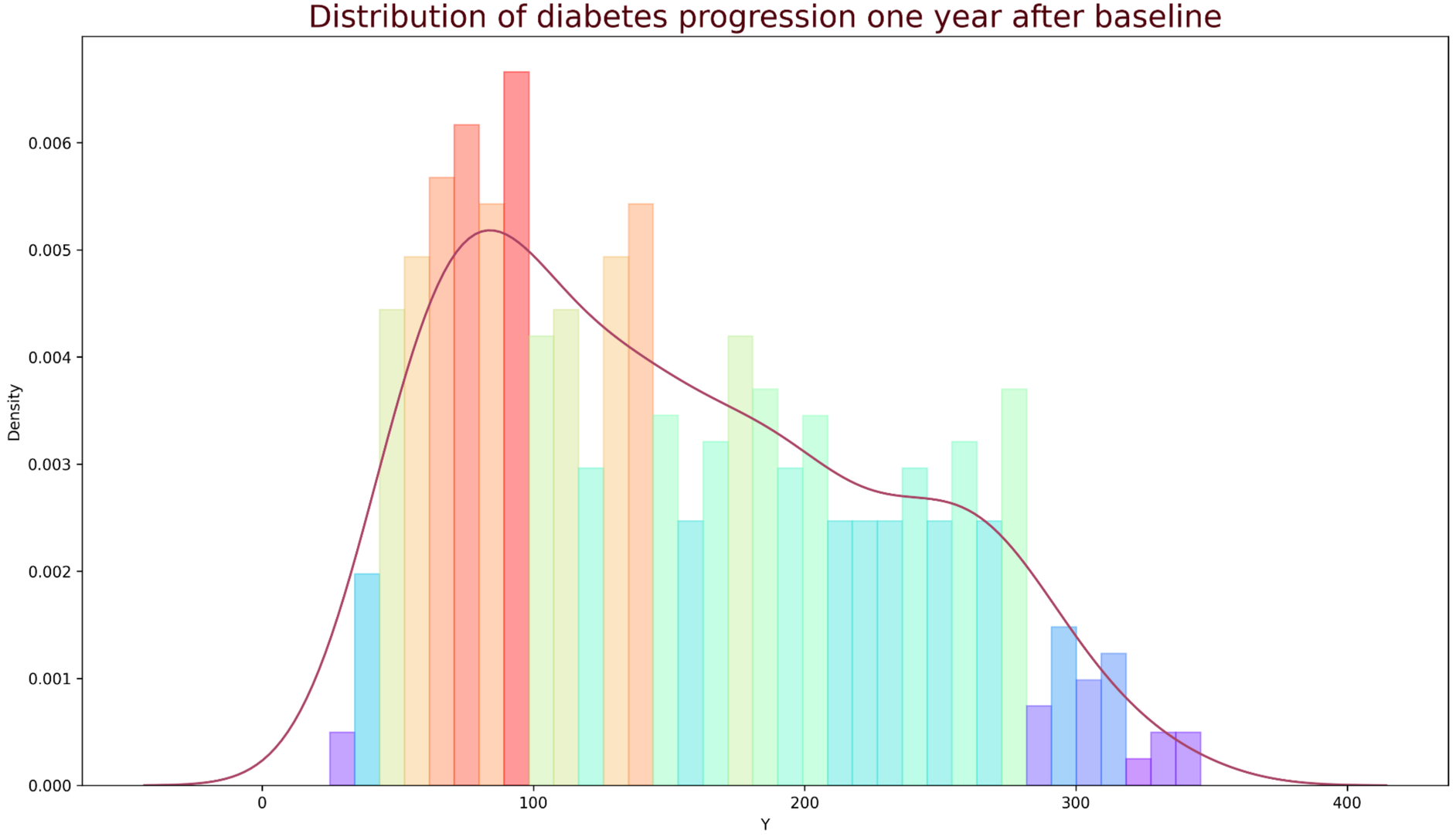


❖ **Nhận xét:**

- Thuộc tính Y tương quan mạnh với thuộc tính **BMI và S5** với hệ số tương quan lần lượt là 0.59 và 0.57.
- Thuộc tính **S1 và S2** có tính tương quan dương rất mạnh với nhau với hệ số tương quan là 0.9, một số thuộc tính cũng có độ tương quan tốt như **S2 và S4** (coef = 0.66), **S4 và S5** (coef = 0.62).
- Thuộc tính **S3** hầu hết có quan hệ nghịch biến (tương quan âm) với các thuộc tính còn lại, trong đó tương quan âm mạnh nhất là với thuộc tính **S4** với hệ số tương quan là -0.74.

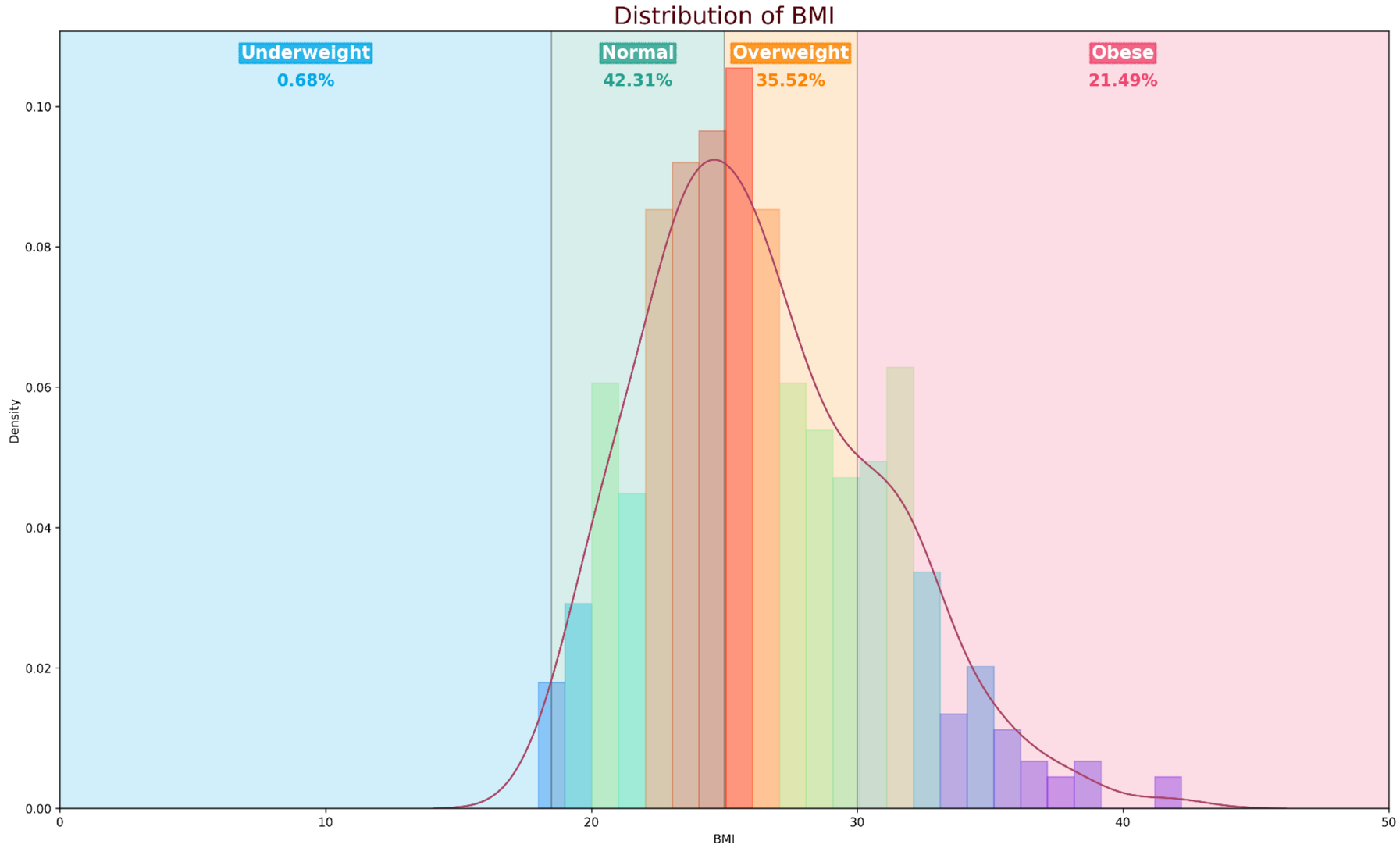
Trực quan bộ dữ liệu

Biểu đồ phân phối giá trị thuộc tính Y



Trực quan bộ dữ liệu

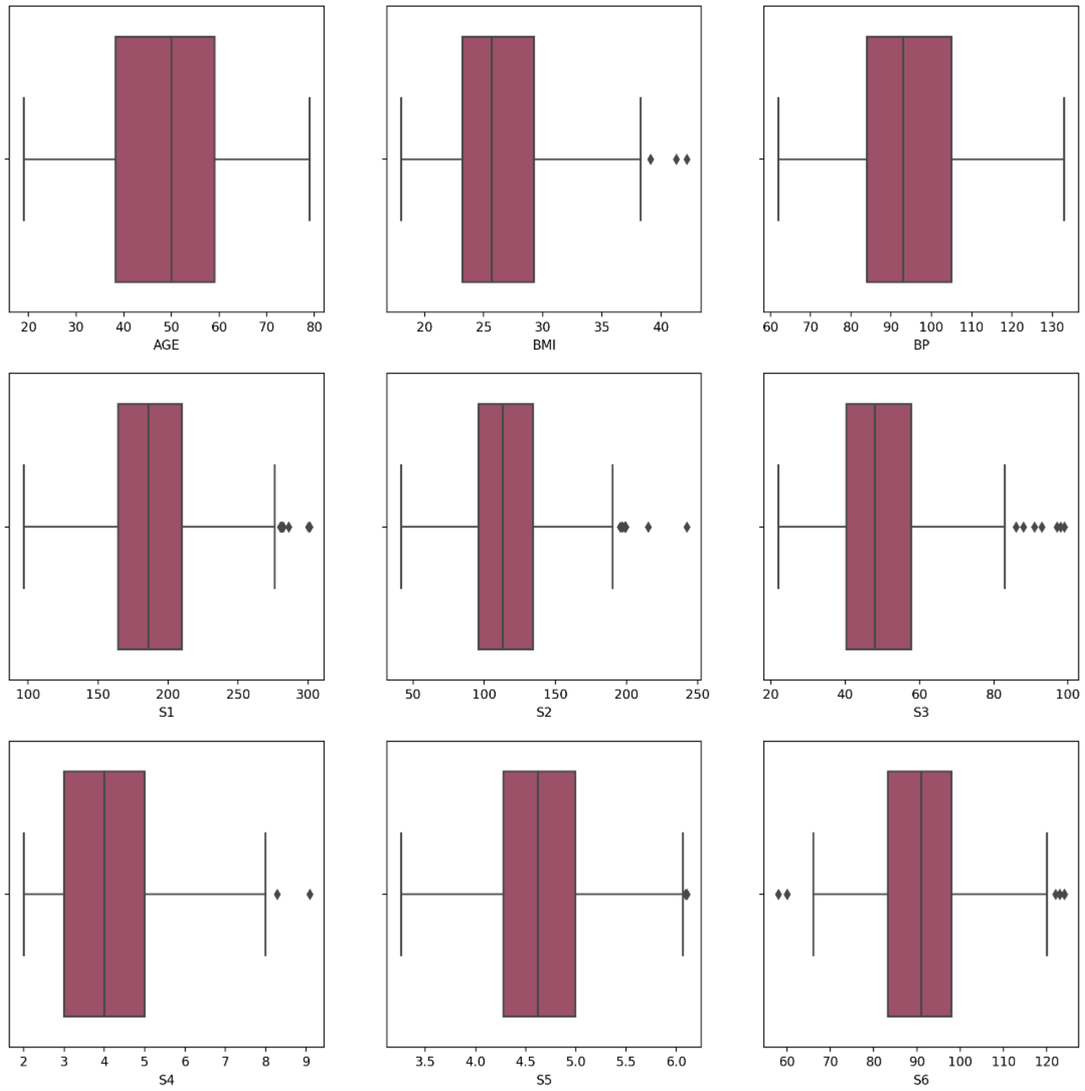
Biểu đồ phân phối giá trị thuộc tính BMI



Tiền xử lý dữ liệu

Xử lý ngoại lệ

Biểu đồ hộp của các thuộc tính định lượng



❖ **Nhận xét:**

- Các thuộc tính **BMI, S1, S2, S3, S4, S5** và **S6** xuất hiện những điểm dữ liệu bất thường.
- Để loại bỏ chúng, dựa vào công thức rút ra từ biểu đồ hộp và tiến hành như sau:

  - Đối với outlier bên trái biểu đồ hộp: Loại bỏ các điểm dữ liệu nhỏ hơn **Q1-1.5*IQR**
  - Đối với outlier bên phải biểu đồ hộp: Loại bỏ các điểm dữ liệu lớn hơn **Q3+1.5*IQR**

  Trong đó:

  ***Q1** là tứ phân vị thứ 25*

  ***Q3** là tứ phân vị thứ 75*

  ***IQR** là hiệu của Q3 và Q1*

Xử lý ngoại lệ

Biểu đồ phân tán của thuộc tính S2 so với Y trước và sau khi tiền xử lý

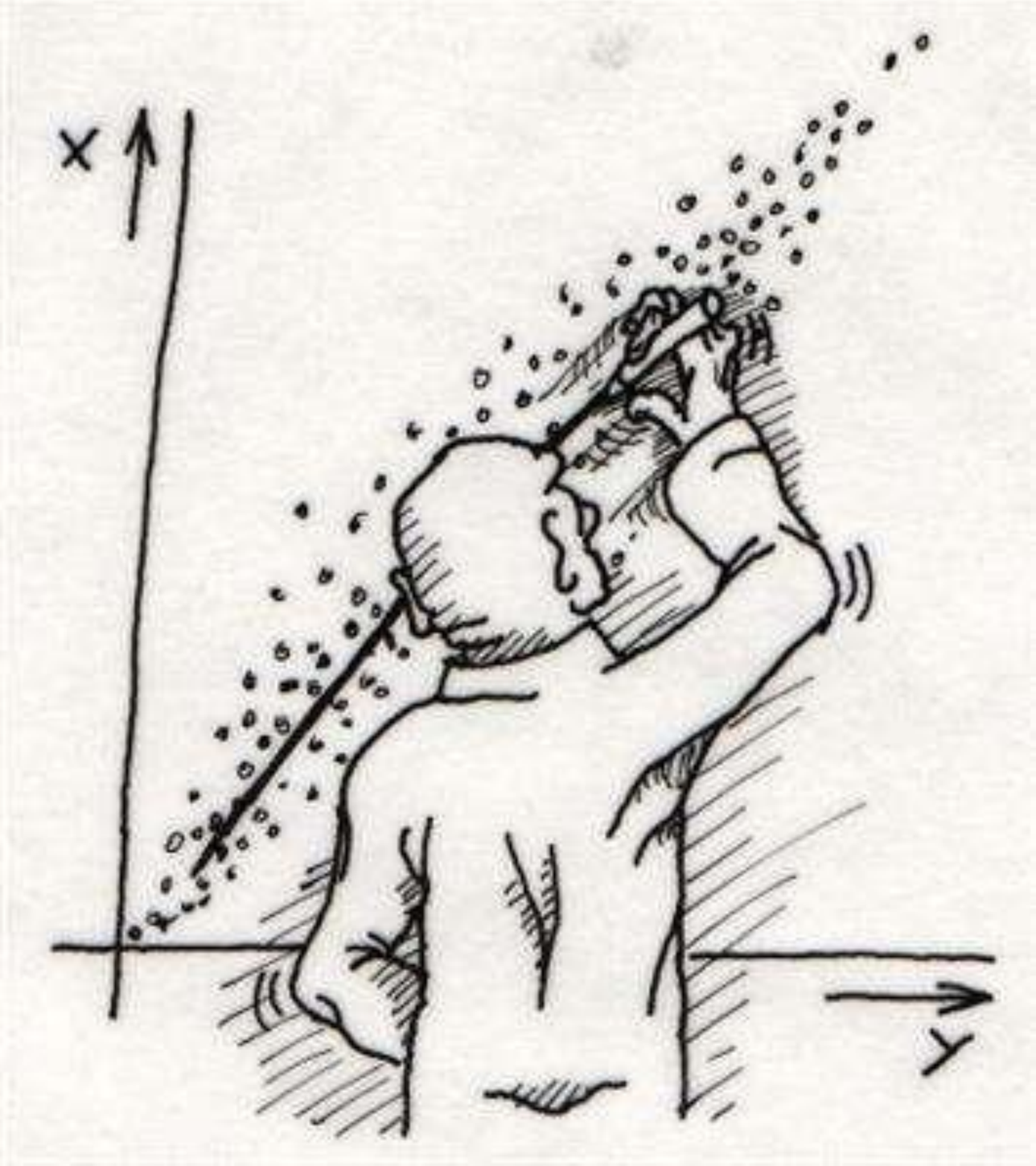
Phân chia tập train và test

Sử dụng hàm **createDataPartition** có sẵn trong **R** để chia tập huấn luyện và tập kiểm thử với tỉ lệ: **80% train – 20% test**.

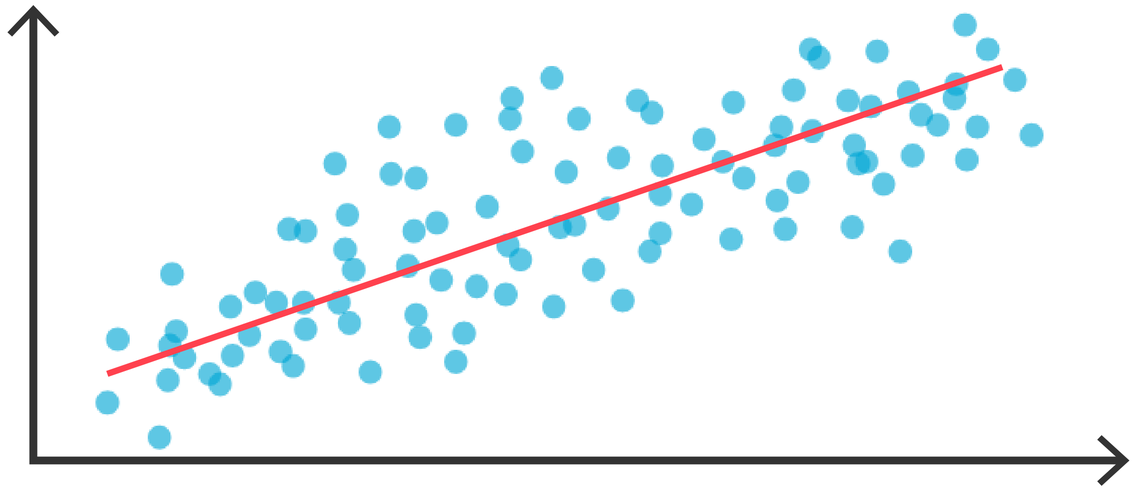
Kích thước của các tập dữ liệu sau khi chia train test

Bộ dữ liệu	Kích thước tập train	Kích thước tập test
Trước khi tiền xử lý	355	87
Sau khi tiền xử lý	327	80

Phân tích hồi quy



Tổng quan về phân tích hồi quy



- Là phương pháp thống kê mạnh mẽ cho phép xem xét mối tương quan giữa các biến số trong việc đo lường.
- Quá trình thực hiện hồi quy cho phép xác định được yếu tố nào đáng quan tâm nhất và những yếu tố nào có thể bỏ qua, hay sự ảnh hưởng, tương tác lẫn nhau của các yếu tố.
- Hai thuật ngữ cần nắm vững trong phân tích hồi quy:
  - **Biến độc lập**
  - **Biến phụ thuộc**

Phân tích ảnh hưởng và tương tác giữa các yếu tố

Trường hợp 1: Khi không xảy ra tương tác giữa các yếu tố

Trường hợp 2: Khi xảy ra tương tác của các yếu tố với SEX

Trường hợp 3: Khi xảy ra tương tác của các yếu tố với S4

Trường hợp 4: Khi xảy ra tương tác giữa các yếu tố còn lại

Phân tích ảnh hưởng và tương tác giữa các yếu tố

Trường hợp 1: Khi không xảy ra tương tác giữa các yếu tố

av <- aov(Y~AGE+SEX+BMI+BP+S1+S2+S3+S4+S5+S6,data=train/train_p)
summary(av)

Trước khi tiền xử lý

```
             Df  Sum Sq Mean Sq F value   Pr(>F)
AGE           1   85855   85855  28.653 1.58e-07 ***
SEX           1     252     252   0.084    0.772
BMI           1  591812  591812 197.514  < 2e-16 ***
BP            1  104605  104605  34.911 8.30e-09 ***
S1            1    5132    5132   1.713    0.191
S2            1    4662    4662   1.556    0.213
S3            1  173465  173465  57.893 2.68e-13 ***
S4            1    1210    1210   0.404    0.526
S5            1   50637   50637  16.900 4.93e-05 ***
S6            1    4458    4458   1.488    0.223
Residuals   344 1030730    2996
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sau khi tiền xử lý

```
             Df Sum Sq Mean Sq F value   Pr(>F)
AGE           1  70622   70622  24.021 1.53e-06 ***
SEX           1      2       2   0.001  0.98169
BMI           1 464953  464953 158.147  < 2e-16 ***
BP            1  91331   91331  31.065 5.35e-08 ***
S1            1   1124    1124   0.382  0.53685
S2            1    746     746   0.254  0.61481
S3            1 223592  223592  76.052  < 2e-16 ***
S4            1    358     358   0.122  0.72744
S5            1  20759   20759   7.061  0.00828 **
S6            1    661     661   0.225  0.63580
Residuals   316 929043    2940
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Khi không xảy ra tương tác, các thuộc tính **AGE, BMI, BP và S5** ảnh hưởng đến tiến triển bệnh tiểu đường với mức ý nghĩa alpha=0.05.
- Các thuộc tính còn lại có p_value > 0.05 cho thấy chúng không ảnh hưởng đến tiến triển bệnh, trong đó SEX và S4 có p_value cao nhất.

  -> Cần xem xét tương tác của chúng với các thuộc tính còn lại.

Phân tích ảnh hưởng và tương tác giữa các yếu tố

Trường hợp 2: Khi xảy ra tương tác của các yếu tố với SEX

```
av <- aov(Y~(AGE+BMI+BP+S1+S2+S3+S5+S6)*SEX,data=train/train_p)
summary(av)
```

Trước khi tiền xử lý

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
AGE	1	85855	85855	29.032	1.34e-07	***
BMI	1	591789	591789	200.115	< 2e-16	***
BP	1	98099	98099	33.172	1.90e-08	***
S1	1	5402	5402	1.827	0.177416	
S2	1	7923	7923	2.679	0.102591	
S3	1	147757	147757	49.965	9.06e-12	***
S5	1	47191	47191	15.958	7.96e-05	***
S6	1	3657	3657	1.237	0.266911	
SEX	1	33132	33132	11.204	0.000909	***
AGE:SEX	1	16983	16983	5.743	0.017101	*
BMI:SEX	1	9617	9617	3.252	0.072228	.
BP:SEX	1	1313	1313	0.444	0.505624	
S1:SEX	1	106	106	0.036	0.850020	
S2:SEX	1	6128	6128	2.072	0.150940	
S3:SEX	1	74	74	0.025	0.874516	
S5:SEX	1	776	776	0.262	0.608866	
S6:SEX	1	425	425	0.144	0.704953	
Residuals	337	996590	2957			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sau khi tiền xử lý

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
AGE	1	70622	70622	24.333	1.33e-06	***
BMI	1	464743	464743	160.132	< 2e-16	***
BP	1	86284	86284	29.730	1.02e-07	***
S1	1	1359	1359	0.468	0.494335	
S2	1	2053	2053	0.707	0.401000	
S3	1	192796	192796	66.430	9.13e-15	***
S5	1	22091	22091	7.612	0.006145	**
S6	1	369	369	0.127	0.721790	
SEX	1	32870	32870	11.326	0.000861	***
AGE:SEX	1	18617	18617	6.415	0.011813	*
BMI:SEX	1	6800	6800	2.343	0.126872	
BP:SEX	1	519	519	0.179	0.672724	
S1:SEX	1	614	614	0.212	0.645764	
S2:SEX	1	1687	1687	0.581	0.446379	
S3:SEX	1	2718	2718	0.936	0.333951	
S5:SEX	1	157	157	0.054	0.816196	
S6:SEX	1	2096	2096	0.722	0.396099	
Residuals	309	896796	2902			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

- Thuộc tính SEX cho thấy ảnh hưởng đến tiến triển bệnh với mức ý nghĩa alpha = 0.05.
- Tương tác giữa AGE với SEX cũng có ảnh hưởng.

Phân tích ảnh hưởng và tương tác giữa các yếu tố

Trường hợp 3: Khi xảy ra tương tác của các yếu tố với S4

av <- aov(Y~(AGE+BMI+BP+S1+S2+S3+S5+S6)*S4,data=train/train_p)
summary(av)

Trước khi tiền xử lý

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
AGE	1	85855	85855	27.968	2.22e-07	***
BMI	1	591789	591789	192.779	< 2e-16	***
BP	1	98099	98099	31.956	3.36e-08	***
S1	1	5402	5402	1.760	0.185547	
S2	1	7923	7923	2.581	0.109082	
S3	1	147757	147757	48.133	2.05e-11	***
S5	1	47191	47191	15.373	0.000107	***
S6	1	3657	3657	1.191	0.275845	
S4	1	599	599	0.195	0.659068	
AGE:S4	1	5122	5122	1.669	0.197321	
BMI:S4	1	10028	10028	3.267	0.071591	.
BP:S4	1	169	169	0.055	0.814705	
S1:S4	1	1218	1218	0.397	0.529180	
S2:S4	1	386	386	0.126	0.723111	
S3:S4	1	182	182	0.059	0.807668	
S5:S4	1	3702	3702	1.206	0.272916	
S6:S4	1	9221	9221	3.004	0.083984	.
Residuals	337	1034516	3070			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sau khi tiền xử lý

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
AGE	1	70622	70622	23.338	2.14e-06	***
BMI	1	464743	464743	153.583	< 2e-16	***
BP	1	86284	86284	28.514	1.81e-07	***
S1	1	1359	1359	0.449	0.50329	
S2	1	2053	2053	0.678	0.41079	
S3	1	192796	192796	63.713	2.85e-14	***
S5	1	22091	22091	7.301	0.00727	**
S6	1	369	369	0.122	0.72731	
S4	1	345	345	0.114	0.73595	
AGE:S4	1	6128	6128	2.025	0.15574	
BMI:S4	1	5028	5028	1.662	0.19836	
BP:S4	1	268	268	0.088	0.76637	
S1:S4	1	14	14	0.005	0.94559	
S2:S4	1	10	10	0.003	0.95391	
S3:S4	1	1404	1404	0.464	0.49630	
S5:S4	1	4644	4644	1.535	0.21635	
S6:S4	1	9999	9999	3.304	0.07007	.
Residuals	309	935035	3026			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

- Thuộc tính S4 và các tương tác với nó không cho thấy ảnh hưởng đến tiến triển bệnh với mức ý nghĩa 0.05.

Phân tích ảnh hưởng và tương tác giữa các yếu tố

Trường hợp 4: Khi xảy ra tương tác giữa các yếu tố

av <- aov(Y~AGE*BMI*BP*S1*S2*S3*S5*S6+AGE*SEX,data=train/train_p)
summary(av)

Trước khi tiền xử lý

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
AGE	1	85855	85855	28.663	5.59e-07	***
BMI	1	591789	591789	197.571	< 2e-16	***
BP	1	98099	98099	32.751	1.12e-07	***
S1	1	5402	5402	1.804	0.182355	
S2	1	7923	7923	2.645	0.107038	
S3	1	147757	147757	49.329	2.77e-10	***
S5	1	47191	47191	15.755	0.000137	***
S6	1	3657	3657	1.221	0.271858	
SEX	1	33132	33132	11.061	0.001238	**
AGE:BMI	1	9377	9377	3.131	0.079915	.
AGE:BP	1	11912	11912	3.977	0.048883	*
BMI:BP	1	21859	21859	7.298	0.008123	**
AGE:S1	1	2	2	0.001	0.979064	
BMI:S1	1	10422	10422	3.479	0.065098	.
BP:S1	1	208	208	0.069	0.792619	
AGE:S2	1	8377	8377	2.797	0.097623	.
BMI:S2	1	523	523	0.175	0.676877	
BP:S2	1	85	85	0.028	0.866446	
S1:S2	1	469	469	0.157	0.693186	
AGE:S3	1	1185	1185	0.396	0.530743	

...

Sau khi tiền xử lý

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
AGE	1	70622	70622	23.261	7.81e-06	***
BMI	1	464743	464743	153.076	< 2e-16	***
BP	1	86284	86284	28.420	1.10e-06	***
S1	1	1359	1359	0.448	0.50567	
S2	1	2053	2053	0.676	0.41368	
S3	1	192796	192796	63.503	1.91e-11	***
S5	1	22091	22091	7.276	0.00872	**
S6	1	369	369	0.121	0.72853	
SEX	1	32870	32870	10.827	0.00156	**
AGE:BMI	1	4782	4782	1.575	0.21357	
AGE:BP	1	147	147	0.048	0.82652	
BMI:BP	1	21990	21990	7.243	0.00887	**
AGE:S1	1	293	293	0.096	0.75700	
BMI:S1	1	1316	1316	0.433	0.51244	
BP:S1	1	129	129	0.043	0.83726	
AGE:S2	1	875	875	0.288	0.59298	
BMI:S2	1	5232	5232	1.723	0.19351	
BP:S2	1	347	347	0.114	0.73628	
S1:S2	1	159	159	0.052	0.81972	
AGE:S3	1	92	92	0.030	0.86254	

...

- Có sự khác nhau về ảnh hưởng của các yếu tố và các tương tác ở bộ dữ liệu trước và sau khi tiền xử lý
- Ở tập train, thuộc tính S1 không ảnh hưởng đến tiến triển bệnh (p_value=0.182355>0.05) nhưng khi tương tác với BMI và S5 thì ảnh hưởng (p_value=0.015230<0.05)

Xây dựng mô hình hồi quy

Simple Linear Regression

Multiple Linear Regression

Polynomial Regression

Elastic Net Regression – Ridge Regression – Lasso Regression

Xây dựng mô hình hồi quy

Simple Linear Regression

Trên tập huấn luyện trước khi tiền xử lý

```
Call:
lm(formula = Y ~ BMI, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-162.809  -43.569   -7.261   48.156  152.338

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -107.096     20.376  -5.256 2.55e-07 ***
BMI            9.846      0.764  12.887  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62.89 on 353 degrees of freedom
Multiple R-squared:  0.3199,        Adjusted R-squared:  0.318
F-statistic: 166.1 on 1 and 353 DF,  p-value: < 2.2e-16
```

Trên tập huấn luyện sau khi tiền xử lý

```
Call:
lm(formula = Y ~ BMI, data = train_p)

Residuals:
     Min       1Q   Median       3Q      Max
-156.418  -45.526   -7.454   47.302  157.173

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -100.1405    21.8255  -4.588 6.39e-06 ***
BMI            9.4412     0.8184  11.536  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62.74 on 325 degrees of freedom
Multiple R-squared:  0.2905,        Adjusted R-squared:  0.2883
F-statistic: 133.1 on 1 and 325 DF,  p-value: < 2.2e-16
```

Xây dựng mô hình hồi quy

Simple Linear Regression

Trên tập huấn luyện trước khi tiền xử lý

$$\mathbf{Y} = -107.096 + 9.846 * \mathbf{BMI}$$

R-squared = 0.3199
Adjusted R-squared = 0.318
RMSE = 62.7097

Trên tập huấn luyện sau khi tiền xử lý

$$\mathbf{Y} = -100.1405 + 9.4412 * \mathbf{BMI}$$

R-squared = 0.2905
Adjusted R-squared = 0.2883
RMSE = 62.5494

Xây dựng mô hình hồi quy

Multiple Linear Regression

Trên tập huấn luyện trước khi tiền xử lý

```
Call:
lm(formula = Y ~ SEX + BMI + BP + S1 + S2 + S5 + S6, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-152.851  -38.319   -0.824   36.942  149.420

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -320.7058    29.6318 -10.823  < 2e-16 ***
SEX          -21.9590     6.5182  -3.369  0.00084 ***
BMI            4.9575     0.8188   6.055 3.65e-09 ***
BP             1.0507     0.2497   4.208 3.29e-05 ***
S1            -1.0232     0.2523  -4.055 6.19e-05 ***
S2             0.8703     0.2622   3.319  0.00100 **
S5            71.9213     8.7834   8.188 5.12e-15 ***
S6             0.3715     0.2944   1.262  0.20784
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.54 on 347 degrees of freedom
Multiple R-squared:  0.4972,        Adjusted R-squared:  0.4871
F-statistic: 49.03 on 7 and 347 DF,  p-value: < 2.2e-16
```

Trên tập huấn luyện sau khi tiền xử lý

```
Call:
lm(formula = Y ~ SEX + BMI + BP + S1 + S3 + S5, data = train_p)

Residuals:
     Min       1Q   Median       3Q      Max
-149.543  -37.501   -0.557   35.273  141.916

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -215.9629    44.5227  -4.851 1.93e-06 ***
SEX          -22.8206     6.8001  -3.356 0.000886 ***
BMI            4.0890     0.8783   4.655 4.74e-06 ***
BP             1.3396     0.2709   4.945 1.23e-06 ***
S1            -0.1542     0.1181  -1.306 0.192529
S3            -1.1371     0.3262  -3.486 0.000560 ***
S5            54.1603     8.6748   6.243 1.36e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.93 on 320 degrees of freedom
Multiple R-squared:  0.4838,        Adjusted R-squared:  0.4741
F-statistic: 49.98 on 6 and 320 DF,  p-value: < 2.2e-16
```

Xây dựng mô hình hồi quy

Multiple Linear Regression

Trên tập huấn luyện trước khi tiền xử lý

$$\mathbf{Y} = -320.7058 - 21.9590 * SEX + 4.9575 * \mathbf{BMI} + 1.0507 * \mathbf{BP} - 1.0232 * \mathbf{S1} + 0.8703 * \mathbf{S2} + 71.9213 * \mathbf{S5} + 0.3715 * \mathbf{S6}$$

R-squared = 0.4972
Adjusted R-squared = 0.4871
RMSE =53.9186

Trên tập huấn luyện sau khi tiền xử lý

$$\mathbf{Y} = -215.9629 - 22.8206 * \mathbf{SEX} + 4.0890 * \mathbf{BMI} + 1.3396 * \mathbf{BP} - 1.542 * \mathbf{S1} - 1.1371 * \mathbf{S3} + 54.1603 * \mathbf{S5}$$

R-squared = 0.4838
Adjusted R-squared = 0.4741
RMSE = 53.3535

Xây dựng mô hình hồi quy

Polynomial Regression

Trên tập huấn luyện trước khi tiền xử lý

```
Call:
lm(formula = Y ~ AGE + BP + BMI + S3 + SEX + I(AGE^2) + I(S5^2) +
    I(S3^2) + I(AGE * SEX) + I(BMI * BP) + I(S1 * S2 * S3 * S5) +
    I(BMI * S1 * S5) + I(BMI * S2 * S5) + I(AGE * S2 * S3 * S6) +
    I(BMI * S1 * S2 * S3 * S5) + I(AGE * BP * S3 * S5 * S6),
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-141.774  -37.239   -4.278   34.056  142.793

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.822e+02  1.603e+02   3.631 0.000326 ***
AGE                      -4.194e+00  1.497e+00  -2.802 0.005378 **
BP                       -3.800e+00  1.333e+00  -2.851 0.004630 **
BMI                      -9.096e+00  4.660e+00  -1.952 0.051797 .
S3                       -4.029e+00  1.788e+00  -2.254 0.024850 *
SEX                      -7.898e+01  2.348e+01  -3.363 0.000858 ***
I(AGE^2)                  1.928e-02  1.490e-02   1.294 0.196462
I(S5^2)                   6.541e+00  2.762e+00   2.368 0.018431 *
I(S3^2)                   2.170e-02  1.296e-02   1.674 0.095126 .
I(AGE * SEX)              1.103e+00  4.586e-01   2.405 0.016708 *
I(BMI * BP)               1.473e-01  4.599e-02   3.203 0.001491 **
I(S1 * S2 * S3 * S5)     -1.270e-05  9.678e-06  -1.312 0.190352
I(BMI * S1 * S5)         -8.057e-03  3.735e-03  -2.157 0.031720 *
I(BMI * S2 * S5)          7.455e-03  4.012e-03   1.858 0.064044 .
I(AGE * S2 * S3 * S6)    -1.541e-06  1.210e-06  -1.273 0.203986
I(BMI * S1 * S2 * S3 * S5) 6.639e-07  3.585e-07   1.852 0.064908 .
I(AGE * BP * S3 * S5 * S6) 8.688e-07  3.069e-07   2.831 0.004916 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.69 on 338 degrees of freedom
Multiple R-squared:  0.543,        Adjusted R-squared:  0.5213
F-statistic:  25.1 on 16 and 338 DF,  p-value: < 2.2e-16
```

Trên tập huấn luyện sau khi tiền xử lý

```
Call:
lm(formula = Y ~ AGE + BMI + S3 + SEX + S5 + I(BP^2) + I(AGE *
    SEX) + I(BMI * BP) + I(BP * S1 * S3 * S6) + I(AGE * BMI *
    S1 * S3) + I(BMI * S2 * S5 * S6) + I(BP * S1 * S2 * S3 *
    S6), data = train_p)

Residuals:
     Min       1Q   Median       3Q      Max
-142.038  -37.410   -0.199   34.083  133.681

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.947e+02  1.329e+02   2.218 0.027300 *
AGE                       -2.731e+00  1.053e+00  -2.594 0.009927 **
BMI                       -1.548e+01  5.445e+00  -2.842 0.004776 **
S3                        -2.081e+00  1.026e+00  -2.028 0.043409 *
SEX                       -9.272e+01  2.422e+01  -3.829 0.000155 ***
S5                         4.012e+01  1.002e+01   4.005 7.76e-05 ***
I(BP^2)                   -1.931e-02  8.467e-03  -2.281 0.023212 *
I(AGE * SEX)               1.431e+00  4.810e-01   2.976 0.003147 **
I(BMI * BP)                1.841e-01  5.534e-02   3.326 0.000985 ***
I(BP * S1 * S3 * S6)       6.006e-07  5.372e-07   1.118 0.264405
I(AGE * BMI * S1 * S3)     2.929e-06  2.529e-06   1.158 0.247655
I(BMI * S2 * S5 * S6)      2.103e-05  1.936e-05   1.086 0.278230
I(BP * S1 * S2 * S3 * S6) -5.426e-09  3.016e-09  -1.799 0.072951 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.78 on 314 degrees of freedom
Multiple R-squared:  0.5149,        Adjusted R-squared:  0.4964
F-statistic: 27.78 on 12 and 314 DF,  p-value: < 2.2e-16
```

Xây dựng mô hình hồi quy

Polynomial Regression

Trên tập huấn luyện trước khi tiền xử lý

$$\mathbf{Y} = 582.2 - 4.194 * \mathbf{AGE} - 3.18 * \mathbf{BP} - 9.096 * \mathbf{BMI} - 4.029 * \mathbf{S3} - 78.98 * \mathbf{SEX} + (1.928e-02) * \mathbf{AGE}^2 + 6.541 * \mathbf{S5}^2 + (2.170e-02) * \mathbf{S3}^2 + 1.103 * \mathbf{AGE*SEX} + (1.473e-01) * \mathbf{BMI*BP} + (-1.270e-05) * \mathbf{S1*S2*S3*S5} + (-8.057e-03) * \mathbf{BMI*S1*S5} + (7.455e-03) * \mathbf{BMI*S2*S5} + (-1.541e-06) * \mathbf{AGE*S2*S3*S6} + (6.639e-07) * \mathbf{BMI*S1*S2*S3*S5} + (8.688e-07) * \mathbf{AGE*BP*S3*S5*S6}$$

R-squared = 0.543
Adjusted R-squared = 0.5213
RMSE = 51.4093

Trên tập huấn luyện sau khi tiền xử lý

$$\mathbf{Y} = 294.7 - 2.731 * \mathbf{AGE} - 15.48 * \mathbf{BMI} - 2.081 * \mathbf{S3} - 92.72 * \mathbf{SEX} + 40.12 * \mathbf{S5} + (-1.931e-02) * \mathbf{BP}^2 + 1.431 * \mathbf{AGE*SEX} + (1.841e-01) * \mathbf{BM*BP} + (6.006e-07) * \mathbf{BP*S1*S3*S6} + (2.929e-06) * \mathbf{AGE*BMI*S1*S3} + (2.103e-05) * \mathbf{BMI*S2*S5*S6} + (-5.426e-09) * \mathbf{BP*S1*S2*S3*S6}$$

R-squared = 0.5149
Adjusted R-squared = 0.4964
RMSE =51.7186

Xây dựng mô hình hồi quy

Ridge Regression

Trên tập huấn luyện trước khi tiền xử lý

```
11 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept) -148.67220469
AGE            0.12349837
SEX           -8.51972205
BMI            3.27742361
BP             0.72694871
S1             0.02000633
S2            -0.03983784
S3            -0.54474454
S4             4.07225690
S5            25.39873274
S6             0.49832477
```

Lambda tốt nhất: 63.87225
R-squared = 0.4763
Adjusted R-squared = 0.4748
RMSE = 56.5607

Trên tập huấn luyện sau khi tiền xử lý

```
11 x 1 sparse Matrix of class "dgCMatrix"
                       s1
(Intercept) -1.447451e+02
AGE          1.147955e-01
SEX         -9.351097e+00
BMI          3.051928e+00
BP           8.154471e-01
S1           2.035922e-03
S2          -6.766365e-02
S3          -6.811350e-01
S4           4.990239e+00
S5           2.890386e+01
S6           3.586102e-01
```

Lambda tốt nhất: 56.88734
R-squared = 0.4684
Adjusted R-squared = 0.4667
RMSE = 55.5186

Xây dựng mô hình hồi quy

Lasso Regression

Trên tập huấn luyện trước khi tiền xử lý

```
11 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept) -156.9429680
AGE            0.0137974
SEX           -5.9714328
BMI            3.4987943
BP             0.7256636

S1             .
S2             .
S3            -0.4931234
S4             3.0923065
S5            27.6310700
S6             0.4366859
```

Lambda tốt nhất: 8.845581
R-squared = 0.4698
Adjusted R-squared = 0.4683
RMSE = 56.5129

Trên tập huấn luyện sau khi tiền xử lý

```
11 x 1 sparse Matrix of class "dgCMatrix"
                       s1
(Intercept) -156.68332510
AGE            0.00625415
SEX           -8.01131052
BMI            3.28575464
BP             0.84842402
S1             .
S2            -0.01193017
S3            -0.69496170
S4             3.61105763
S5            32.11917544
S6             0.26372880
```

Lambda tốt nhất: 7.87825
R-squared = 0.4592
Adjusted R-squared = 0.4575
RMSE = 55.5805

Xây dựng mô hình hồi quy

Elastic Net Regression (alpha = 0.05)

Trên tập huấn luyện trước khi tiền xử lý

```
11 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept) -198.2166275
AGE            .
SEX            .
BMI            4.7320556
BP             0.5720361
S1             .
S2             .
S3            -0.2404157
S4             .
S5            39.4848485
S6             .
```

Lambda tốt nhất: 48.09499
R-squared = 0.4774
Adjusted R-squared = 0.4759
RMSE = 56.5372

Trên tập huấn luyện sau khi tiền xử lý

```
11 x 1 sparse Matrix of class "dgCMatrix"
                       s1
(Intercept) -156.68332510
AGE            0.00625415
SEX           -8.01131052
BMI            3.28575464
BP             0.84842402
S1             .
S2            -0.01193017
S3            -0.69496170
S4             3.61105763
S5            32.11917544
S6             0.26372880
```

Lambda tốt nhất: 39.03005
R-squared = 0.469
Adjusted R-squared = 0.4674
RMSE = 55.3056

Đánh giá mô hình hồi quy

Kết quả đánh giá mô hình

Kết quả đánh giá **Adjusted R-Squared** của các mô hình trên tập huấn luyện

Tên mô hình	Adjusted R-Squared		RMSE	
	Trước khi tiền xử lý	Sau khi tiền xử lý	Trước khi tiền xử lý	Sau khi tiền xử lý
Simple Linear Regression	0.3180	0.2883	62.7097	62.5494
Multiple Linear Regression	0.4871	0.4741	53.9186	53.3535
Polynomial Regression	0.5213	0.4964	51.4093	51.7186
Ridge Regression	0.4748	0.4667	56.5607	55.5186
Lasso Regression	0.4683	0.4575	56.5129	55.5805
Elastic Net Regression	0.4759	0.4674	56.5372	55.3056

❖ **Nhận xét:**

- Các mô hình trên tập dữ liệu huấn luyện trước khi tiền xử lý có kết quả Adjusted R-Squared tốt hơn so với sau khi tiền xử lý. Còn khi xét độ đo RMSE thì hầu hết mô hình phạm ít lỗi hơn khi sử dụng dữ liệu đã qua xử lý.
- Độ đo đánh giá Adjusted R-Squared cao nhất trước và sau khi tiền xử ý trên tập dữ liệu huấn luyện đều thuộc về mô hình **Polynomial Regression** với kết quả lần lượt là 0.5213 và 0.4964. Điều đó cho thấy mô hình Polynomial Regression có mức độ tương thích tốt hơn so với các mô hình còn lại trên tập dữ liệu huấn luyện.
- Độ đo đánh giá Adjusted R-Squared thấp nhất trước và sau khi tiền xử lí trên tập dữ liệu huấn luyện đều thuộc về mô hình **Simple Linear Regression** với kết quả lần lượt là 0.318 và 0.2883. Điều đó cho thấy mô hình Simple Linear Regression chưa phù hợp trên tập dữ liệu huấn luyện.

Kết quả đánh giá mô hình

Kết quả đánh giá **Adjusted R-squared** của các mô hình trên tập kiểm thử

Tên mô hình	Adjusted R-Squared		RMSE	
	Trước khi tiền xử lý	Sau khi tiền xử lý	Trước khi tiền xử lý	Sau khi tiền xử lý
Simple Linear Regression	0.4334	0.5099	61.1129	59.0436
Multiple Linear Regression	0.5795	0.4998	52.5026	56.5038
Polynomial Regression	0.5499	0.5669	53.9409	52.7150
Ridge Regression	0.5354	0.4833	58.0027	59.6312
Lasso Regression	0.5486	0.4792	56.9988	58.5861
Elastic Net Regression	0.5395	0.4863	57.8680	59.2369

❖ **Nhận xét:**

- Đa số mô hình trên tập dữ liệu kiểm thử trước khi tiền xử lí có kết quả tốt hơn so với sau khi tiền xử lí, riêng mô **hình Simple Linear Regression** và **Polynomial Regression** sau khi tiền xử lí lại có hiệu suất cao hơn so với trước khi tiền xử lí.
- Độ đo đánh giá Adjusted R-Squared cao nhất trên tập dữ liệu kiểm thử trước khi tiền xử lí thuộc về mô hình **Multiple Linear Regression** với kết quả là 0.5795 và sau khi tiền xử lí thuộc về mô hình **Polynomial Regression** với kết quả là 0.5669.
- Độ đo đánh giá Adjusted R-Squared thấp nhất trên tập dữ liệu kiểm thử trước khi tiền xử lí thuộc về mô hình Simple Linear Regression với kết quả là 0.4334 và sau khi tiền xử lí thuộc về mô hình **Lasso Regression** với kết quả là 0.4792.

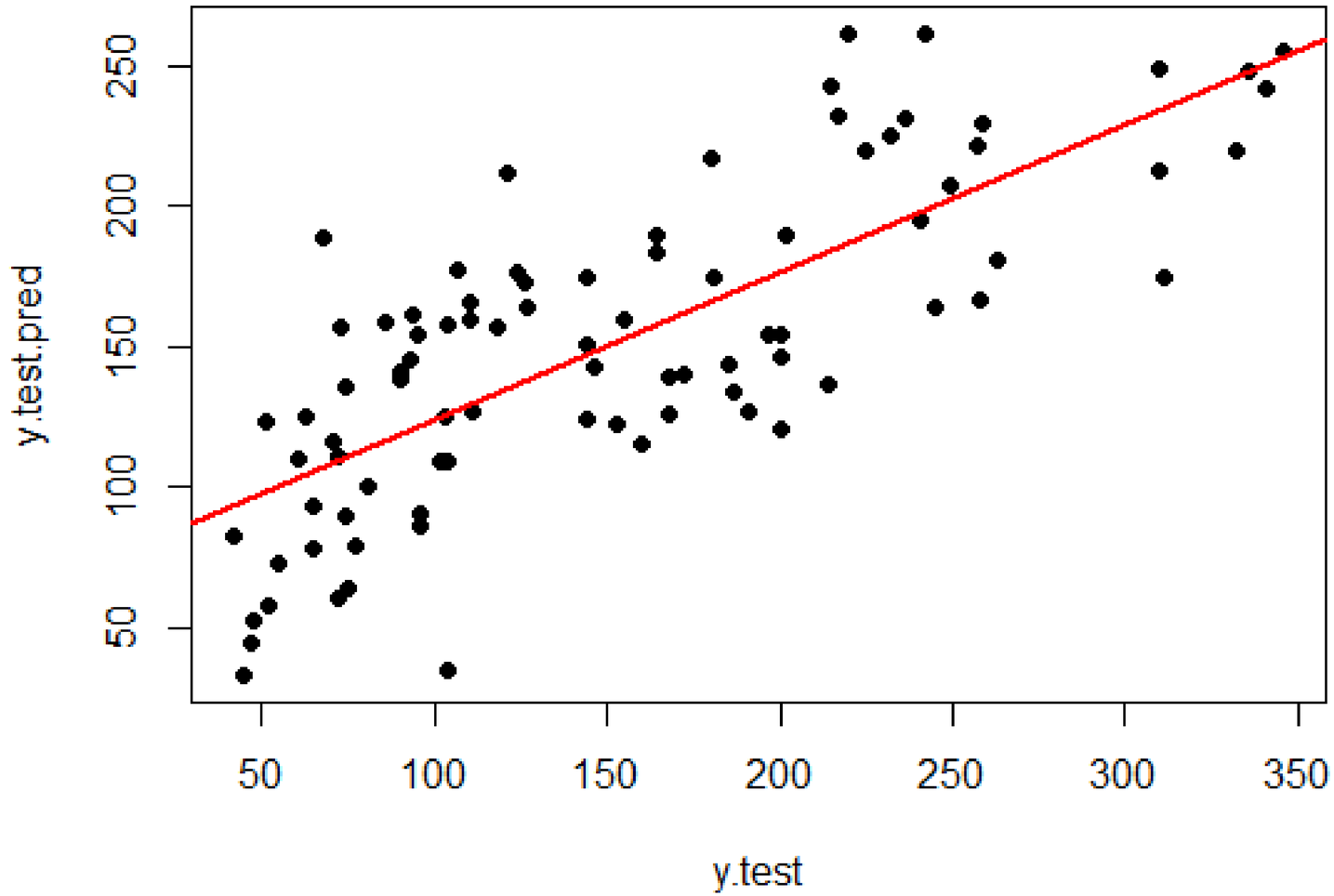
Mô hình tốt nhất

Mô hình đạt hiệu suất tốt nhất trên tập kiểm thử là mô hình **Multiple Linear Regression** trên tập dữ liệu **chưa qua tiền xử lý**.

$$Y = -320.7058 - 21.9590 * \mathbf{SEX} + 4.9575 * \mathbf{BMI} + 1.0507 * \mathbf{BP} - 1.0232 * \mathbf{S1} + 0.8703 * \mathbf{S2} + 71.9213 * \mathbf{S5} + 0.3715 * \mathbf{S6}$$

Mô hình tốt nhất

Biểu đồ thể hiện mối quan hệ giữa giá trị dự đoán và giá trị kiểm thử
Mô hình Multiple Linear Regression



Kết luận

Kết luận

- Kết quả đánh giá Adjusted R-Squared nhìn chung khá thấp do các thuộc tính ít có quan hệ tuyến tính với Y và có nhiều outlier.
- Việc loại bỏ các điểm dữ liệu ngoại lệ có thể giảm hiệu suất của mô hình.
- Có những yếu tố khi sử dụng đơn lẻ thì không có ảnh hưởng đến kết quả, điển hình trong bộ dữ liệu là hai thuộc tính SEX và S1, nhưng khi kết hợp với yếu tố khác có thể ảnh hưởng tới kết quả.
- Không chỉ chú ý tới các yếu tố có ảnh hưởng tới kết quả mà cần phải quan tâm tới tương tác giữa các yếu tố.
- Để mô hình đạt hiệu suất cao thì cần phải liên tục thực nghiệm trên nhiều yếu tố và tương tác khác nhau, cân nhắc xử lý các điểm dữ liệu bất thường,...

Cảm ơn thầy và các bạn đã lắng nghe!