Câu 1

Có một tập dữ liệu này gồm 500 mẫu về 3 loài hoa khác nhau của họ Iris là (Iris setosa, Iris virginica và Iris versicolor). Với mỗi một mẫu hoa người ta thu thập bốn thuộc tính là chiều dài và chiều rộng của đài hoa và cánh hoa với đơn vị centimet. Người ta mong muốn xây dựng một phần mềm trên điện thoại di động. Khi người dùng nhập vào 4 thông tin này thì phần mềm phải dự đoán được đây thuộc họ iris nào. Bạn hãy đề xuất giải thuật phù hợp để xây dựng một mô hình dự đoán này. Hãy giải thích sự lựa chọn của bạn.

a. Giải thuật học có giám sát, giải thuật phân lớp (classification).

b. Giải thuật học có giám sát, giải thuật hồi quy (regression).

c. Giải thuật học không có giám sát, giải thuật gom nhóm, gom cụm (clustering).

d. Tất cả các giải thuật bên trên đều không phù hợp.

Trả lời:

Câu 2

Chúng ta hãy xét một tập dữ liệu số thu được từ một nghiên cứu kiểm tra nhiệt lượng phát sinh trong quá trình đông cứng của xi măng Portland. Nhiệt lượng này được giả định là một hàm của các thành phần hóa học, các biến sau đây đã được đo:

x1: số lượng tricalcium aluminat

x2: số lượng tricalcium silicat

x3: số lượng ferit nhôm tetracalcium

x4: số lượng dicalcium silicat

Y: nhiệt lượng toả ra tính bằng calo trên mỗi gram xi măng.

Bảng 1: Dữ liệu quan sát

i	x_1	x_2	x_3	x_4	*Y*
1	7	6	26	60	78,5
2	1	26	15	52	74,3
3	11	56	8	20	104,3

4	11	31	8	47	87,6
5	7	52	6	33	95,9
6	11	55	9	22	109,2
7	3	71	17	6	102,7
8	1	31	22	44	72,5
9	2	54	18	22	93,8
10	21	47	4	26	115,9
11	1	40	23	34	83,8
12	11	66	9	12	113,3
13	10	68	8	12	109,4

Nguồn: Bài giảng của thầy Nguyễn Văn Tuấn

Các biến x_1, x_2, x_3 và x_4 được tính theo phần trăm trọng lượng của clanhke làm ra xi măng. Người ta mong muốn sử dụng các giải thuật máy học để xây dựng mô hình dự đoán nhiệt lượng toả ra (Y) từ 4 biến x_1, x_2, x_3 và x_4. Bạn hãy đề xuất giải thuật phù hợp để xây dựng một mô hình dự đoán này. Hãy giải thích sự lựa chọn của bạn.

a. Giải thuật học có giám sát, giải thuật phân lớp (classification).

b. Giải thuật học có giám sát, giải thuật hồi quy (regression).

c. Giải thuật học không có giám sát, giải thuật gom nhóm, gom cụm (clustering).

d. Tất cả các giải thuật bên trên đều không phù hợp.

Trả lời:

Câu 3

Có 1 thí nghiệm nghiên cứu ảnh hưởng của các yếu tố như nhiệt độ, thời gian, và thành phần hóa học, ... đến sản lượng CO2. Số liệu của nghiên cứu này có thể tóm lược trong bảng sau:

Id	y	X1	X2	X3	X4	X5	X6	X7
1	36.98	5.1	400	51.37	4.24	1484.83	2227.25	2.06
2	13.74	26.4	400	72.33	30.87	289.94	434.90	1.33
3	10.08	23.8	400	71.44	33.01	320.79	481.19	0.97
4	8.53	46.4	400	79.15	44.61	164.76	247.14	0.62
5	36.42	7.0	450	80.47	33.84	1097.26	1645.89	0.22
6	26.59	12.6	450	89.90	41.26	605.06	907.59	0.76
7	19.07	18.9	450	91.48	41.88	405.37	608.05	1.71
8	5.96	30.2	450	98.60	70.79	253.70	380.55	3.93
9	15.52	53.8	450	98.05	66.82	142.27	213.40	1.97
10	56.61	5.6	400	55.69	8.92	1362.24	2043.36	5.08
11	26.72	15.1	400	66.29	17.98	507.65	761.48	0.60
12	20.80	20.3	400	58.94	17.79	377.60	566.40	0.90
13	6.99	48.4	400	74.74	33.94	158.05	237.08	0.63
14	45.93	5.8	425	63.71	11.95	130.66	1961.49	2.04
15	43.09	11.2	425	67.14	14.73	682.59	1023.89	1.57
16	15.79	27.9	425	77.65	34.49	274.20	411.30	2.38
17	21.60	5.1	450	67.22	14.48	1496.51	2244.77	0.32
18	35.19	11.7	450	81.48	29.69	652.43	978.64	0.44
19	26.14	16.7	450	83.88	26.33	458.42	687.62	8.82
20	8.60	24.8	450	89.38	37.98	312.25	468.38	0.02
21	11.63	24.9	450	79.77	25.66	307.08	460.62	1.72
22	9.59	39.5	450	87.93	22.36	193.61	290.42	1.88
23	4.42	29.0	450	79.50	31.52	155.96	233.95	1.43
24	38.89	5.5	460	72.73	17.86	1392.08	2088.12	1.35
25	11.19	11.5	450	77.88	25.20	663.09	994.63	1.61
26	75.62	5.2	470	75.50	8.66	1464.11	2196.17	4.78
27	36.03	10.6	470	83.15	22.39	720.07	1080.11	5.88

(Nguồn: Bài giảng của thầy Nguyễn Văn Tuấn)

Chú thích: y = sản lượng CO2;

X1 = thời gian (phút);
X2 = nhiệt độ (C);
X3 = phần trăm hòa tan;
X4 = lượng dầu (g/100g);

X5 = lượng than đá;
X6 = tổng số lượng hòa tan;
X7 = số hydrogen tiêu thụ.

Bạn chọn giải thuật nào sau đâu xây dựng các mô hình dự báo sản lượng CO2 (giá trị Y trong bảng bên trên) từ 7 thuộc tính (từ X1 đến X7). Hãy giải thích cho sự lựa chọn của bạn.

a. Một trong các giải thuật hồi quy (regression).

b. Một trong các giải thuật phân lớp (classfication).

c. Một trong các giải thuật gom nhóm (clustering).

d. Các giải thuật bên trên đều không phù hợp

Trả lời:

Câu 4

Chúng ta xem xét ví dụ sau, trong đó các nhà nghiên cứu đo lường độ cholestrol trong máu của 18 đối tượng nam. Tỉ trong cơ thể (body mass index) cũng được ước tính cho mỗi đối tượng bằng thức tính BMI là lấy trọng lượng (tính bằng kg) chia cho chiều cao bình phương (m2). Kết quả đo lường như sau

Bảng . Độ tuổi, tỉ trọng cơ thể và cholesterol

Mã số ID (id)	Độ tuổi (age)	BMI (bmi)	Cholesterol (chol)
1	46	25.4	3.5
2	20	20.6	1.9
3	52	26.2	4.0
4	30	22.6	2.6
5	57	25.4	4.5
6	25	23.1	3.0
7	28	22.7	2.9
8	36	24.9	3.8
9	22	19.8	2.1
10	43	25.3	3.8
11	57	23.2	4.1
12	33	21.8	3.0
13	22	20.9	2.5
14	63	26.7	4.6
15	40	26.4	3.2
16	48	21.2	4.2
17	28	21.2	2.3
18	49	22.8	4.0

(Nguồn: Bài giảng của thầy Nguyễn Văn Tuấn)

Người ta muốn xây dựng mô hình ước đoán độ cholesterol dựa vào hai biến số độ tuổi và bmi. Bạn hãy đề xuất giải thuật phù hợp. Hãy giải thích cho sự lựa chọn của bạn.

a. Một trong các giải thuật hồi quy (regression).

b. Một trong các giải thuật phân lớp (classfication).

c. Một trong các giải thuật gom nhóm (clustering).

d. Các giải thuật bên trên đều không phù hợp

Trả lời:

Câu 5.

Trong nghiên cứu (Jolicoeu, 1960), người ta đo chiều dài, chiều rộng và chiều cao của 25 con rùa. Dữ liệu thu thập được như sau:

length	width	height
98	81	38
103	84	38
103	86	42
105	86	42
109	88	44
123	92	50
123	95	46
133	99	51
133	102	51

…

Nguồn: Jolicoeur, P. and Mosimann, J.E., 1960. Size and shape variation in the painted turtle. A principal component analysis. Growth, 24(4), pp.339-354.

Bạn hãy xây dựng phần mềm cho phép nhà nghiên cứu phân tích nhóm các con rùa trên. Bạn hãy đề xuất giải thuật phù hợp. Hãy giải thích cho sự lựa chọn của bạn.

a. Một trong các giải thuật hồi quy (regression).

b. Một trong các giải thuật phân lớp (classfication).

c. Một trong các giải thuật gom nhóm (clustering).

d. Các giải thuật bên trên đều không phù hợp

Trả lời: