Họ và tên	**Kiều Công Hậu**
MSSV	**18127259**
Lớp	18CLC1

TOÁN ỨNG DỤNG VÀ THỐNG KÊ

Đồ án 3: Linear regression

1. Các chức năng đã hoàn thành:

STT	Yêu cầu	Hoàn thành
1	Sử dụng toàn bộ 11 đặc trưng đề bài cung cấp.	100%
2	Sử dụng duy nhất 1 đặc trưng cho kết quả tốt nhất.	100%
3	Xây dựng một mô hình của riêng bạn cho kết quả tốt nhất.	100%

2. Ý tưởng thực hiện, mô tả các hàm và kết quả của từng yêu cầu:

a. Sử dụng toàn bộ 11 đặc trưng đề bài cung cấp.

- Ý tưởng: tìm x bằng công thức $x = A^{\dagger}b$ với A là ma trận được lấy từ bộ dữ liệu bỏ đi cột Quality và b là vector của cột Quality.
- Giải thích các hàm:
  - calc_x_hat: Tính x mũ cho mô hình hồi quy tuyến tính dựa vào bộ dữ liệu. (A @ x = b)
  - linear_regression: Kết quả của phương pháp hồi quy tuyến tính áp dụng trên 1 bộ dữ liệu. Hàm này chỉ đơn giản là hàm in ra các giá trị của vector x_hat đã được tính từ hàm calc_x_hat nêu trên.
- Kết quả:

	x	Tính chất	Giá trị
0	x1	fixed acidity	0.005925
1	x2	volatile acidity	-1.108038
2	x3	citric acid	-0.263046
3	x4	residual sugar	0.015322
4	x5	chlorides	-1.730503
5	x6	free sulfur dioxide	0.003801
6	x7	total sulfur dioxide	-0.003899
7	x8	density	4.338588
8	x9	pH	-0.458535
9	x10	sulphates	0.729719
10	x11	alcohol	0.308859

b. Sử dụng duy nhất 1 đặc trưng cho kết quả tốt nhất.

- Ý tưởng: Sử dụng phương pháp Cross Validation để tính sai số trung bình cho 11 bộ dữ liệu, mỗi bộ dữ liệu chỉ có 1 tính chất duy nhất. Bộ dữ liệu tốt nhất là bộ dữ liệu có sai số trung bình thấp nhất trong 11 bộ. Ứng với mỗi bộ dữ liệu, ta sẽ chia bộ đó thành 4 phần bằng nhau (tương đối), tạm gọi là bộ A, bộ B, bộ C và bộ D. Trong trường hợp cụ thể này, bộ dữ liệu có 1199 dòng dữ liệu nên bộ A, B, C, mỗi bộ có 300 dòng, riêng bộ D có 299 dòng. Ta chọn ra 3 bộ để train, 1 bộ còn lại để test. Từ bộ test này kết hợp với mô hình hồi quy tuyến tính được xây dựng từ 3 bộ train, ta có thể dễ dàng tính được sai số trung bình. Tuy nhiên, để chính xác hơn, ta sẽ thay phiên bộ dữ liệu đóng vai trò bộ test, vậy nên ứng với mỗi bộ dữ liệu, ta sẽ tính sai số 4 lần, đó là tính Sai số A, Sai số B, Sai số C và Sai số D (chú thích: Sai số A là sai số được tính từ bộ dự liệu có A là bộ test, 3 bộ còn lại là bộ train,…). Có được 4 giá trị sai số trên, ta lấy trung bình (cộng lại chia 4) để ra sai số trung bình của mô hình ứng với bộ dữ liệu mà ta đang xét. Sau khi tính được toàn bộ 11 sai số trung bình, ta bắt đầu xếp hạng để chọn ra bộ dữ liệu tốt nhất (có sai số bé nhất) để lập mô hình hồi quy tuyến tính.

- Giải thích hàm:
  - calc_residual: Tính sai số từ 1 cặp bộ dữ liệu train và test.
  - calc_avg_residual: Tính sai số trung bình của 1 bộ dữ liệu theo phương pháp Cross Validation. Cụ thể hàm này sẽ gọi hàm calc_residual 4 lần để cộng lại rồi chia 4 để ra sai số trung bình. Hàm này trả về sai số trung bình của bộ dữ liệu này và danh sách 4 sai số A, B, C và D.
  - calc_feature_avg_residuals: Tính các sai số trung bình của các bộ dữ liệu (mỗi bộ dữ liệu ứng với duy nhất 1 đặc trưng). Cụ thể hàm này sẽ quét vòng for qua 11 bộ dữ liệu, mỗi bộ chỉ có 1 tính chất và gọi hàm cal_avg_residual để tính sai số trung bình của bộ dữ liệu đang duyệt.
  - ranking_table: Xếp hạng so sánh sai số của mỗi bộ dữ liệu ứng với mỗi tính chất để in ra màn hình dưới dạng bảng.
  - build_best_data: Lọc bộ dữ liệu gốc chỉ còn 1 tính chất được cho là tốt nhất (cho kết quả sai số bé nhất).
- Kết quả:
  - Bảng xếp hạng:

	Tính chất	Sai số A	Sai số B	Sai số C	Sai số D	Sai số trung bình	Xếp hạng
0	alcohol	0.556146	0.525760	0.596566	0.512212	0.547671	1
1	density	0.782727	0.696102	0.685753	0.694613	0.714799	2
2	pH	0.791555	0.722449	0.736420	0.720151	0.742644	3
3	fixed acidity	1.221617	0.856488	1.239303	0.991560	1.077242	4
4	sulphates	1.200110	1.253365	0.935362	1.004242	1.098270	5
5	volatile acidity	2.296810	1.618290	1.841731	1.700559	1.864348	6
6	residual sugar	2.545658	1.988977	1.894290	1.628064	2.014247	7
7	chlorides	2.536739	1.932540	2.099208	1.869860	2.109587	8
8	citric acid	2.488746	2.948598	2.370226	2.786764	2.648583	9
9	free sulfur dioxide	3.255725	2.518466	3.123854	2.476097	2.843535	10
10	total sulfur dioxide	3.647015	2.798878	3.230203	2.880968	3.139266	11

- Mô hình:

	x	Tính chất	Giá trị
0	x1	alcohol	0.543706

c. Xây dựng một mô hình của riêng bạn cho kết quả tốt nhất.

- Ý tưởng: Thêm 1 tính chất nữa vào bộ dữ liệu gốc, tính chất này là tổng của 2 tính chất có sai số bé nhất. Lý do là ta muốn giảm sai số của mô hình nên buộc phải tăng sự ảnh hưởng của các tính chất tốt nhất vào bộ dữ liệu. Cụ thể, 2 tính chất tốt nhất là alcohol và density, vậy bộ dữ liệu mới sẽ có 12 tính chất, với hy vọng sai số sẽ giảm.
- Giải thích hàm:
  - build_my_data: Xây dựng lại bộ dữ liệu mới tốt hơn dựa trên bộ dữ liệu gốc bằng cách thêm 1 tính chất mới là tổng của 2 tính chất tốt nhất được xếp hạng từ câu b.
  - compare: So sánh sai số của mô hình câu a (11 tính chất) và sai số của mô hình mới (12 tính chất).
- Kết quả: Quả thật sai số của mô hình mới này có giảm đi một chút, dù không đáng kể nhưng cho thấy ta đã đúng hướng về mặt ý tưởng cải thiện: "tăng sự ảnh hưởng của các tính chất tốt nhất vào bộ dữ liệu". Cụ thể:
  - So sánh:

```
* 0.5150558048747267 là sai số của mô hình hồi quy tuyến tính với 11 tính chất.
* 0.5150558048747231 là sai số của mô hình hồi quy tuyến tính mới.

Sai số của mô hình mới bé hơn sai số của mô hình với 11 tính chất.
(3.552713678800501e-15 là chênh lệch giữa 2 sai số này)
```

  - Mô hình:

	x	Tính chất	Giá trị
0	x1	Tính chất mới	1.549149
1	x2	fixed acidity	0.005925
2	x3	volatile acidity	-1.108038
3	x4	citric acid	-0.263046
4	x5	residual sugar	0.015322
5	x6	chlorides	-1.730503
6	x7	free sulfur dioxide	0.003801
7	x8	total sulfur dioxide	-0.003899
8	x9	density	2.789439
9	x10	pH	-0.458535
10	x11	sulphates	0.729719
11	x12	alcohol	-1.240290

HẾT