TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐẠI HỌC QUỐC GIA TP.HCM
KHOA CÔNG NGHỆ THÔNG TIN – CHẤT LƯỢNG CAO

LAB 4 – PROJECT 3 – LINEAR REGRESSION

MÔN TOÁN ỨNG DỤNG VÀ THỐNG KÊ CHO CÔNG NGHỆ THÔNG TIN

Lớp: *21CLC08*
Giảng viên hướng dẫn: *Phan Thị Phương Uyên*

Sinh viên thực hiện:

- 21127386 – Nguyễn Thị Cẩm Nhung

Thành phố Hồ Chí Minh, Tháng 8/ 2023

MỤC LỤC

I. Các thư viện sử dụng:

Thư viện	Lý do sử dụng
panda	Nhập dữ liệu đầu vào với dạng DataFrame
numpy	Tính toán các phép toán ma trận
sklearn	Sử dụng các hàm có sẵn như LinearRegression, KFold, tính MAE

II. Các hàm sử dụng và mô tả hàm:

1. Yêu cầu 1a:

- *Đầu vào:* 11 đặc trưng đề bài cung cấp.
- *Mô tả hàm tìm ra hệ số:* Đọc các dữ liệu đầu vào từ data cho sẵn vào các biến *X_train1b* và *Y_train1b*, tạo *model1a* là một mô hình hồi quy tuyến tính bình phương nhỏ nhất, sử dụng lớp *LinearRegression()* thuộc thư viện sklearn, sau đó dùng phương thức *fit()* thuộc lớp trên để huấn luyện cho yêu cầu, phương thức *fit()* sẽ dựa trên phương pháp bình phương nhỏ nhất (Least Square) để tìm ra các hệ số cho *model1a*.

Bộ thư viện Scikit-learn rất là đồ sộ, do chúng ta đã tìm hiểu về thuật toán Linear Regression nên chúng ta chỉ import phần liên quan đến thuật toán này mà thôi. Để thực hiện thuật toán Hồi quy tuyến tính trên các dữ liệu X, y đã được bóc tách và mô phỏng trong bài trước, chúng ta thực hiện đoạn code trong Python:

```
regression = LinearRegression()
regression.fit(X, y)
```

Biến regression chứa một đối tượng LinearRegression trong bộ thư viện Scikit-learn và tiếp theo là thực hiện phương thức fit() trên đối tượng này. Fit() thực hiện tính toán tối ưu hóa các tham số θ0 và θ1, phương thức này trả về một model. Sau khi thực hiện phương thức này, chúng ta đã có một đối tượng chứa đầy đủ thông tin kết quả, ví dụ chúng ta cần lấy về giá trị của các tham số θ0 và θ1:

```
# theta_1
regression.coef_

# theta_0
regression.intercept_
```

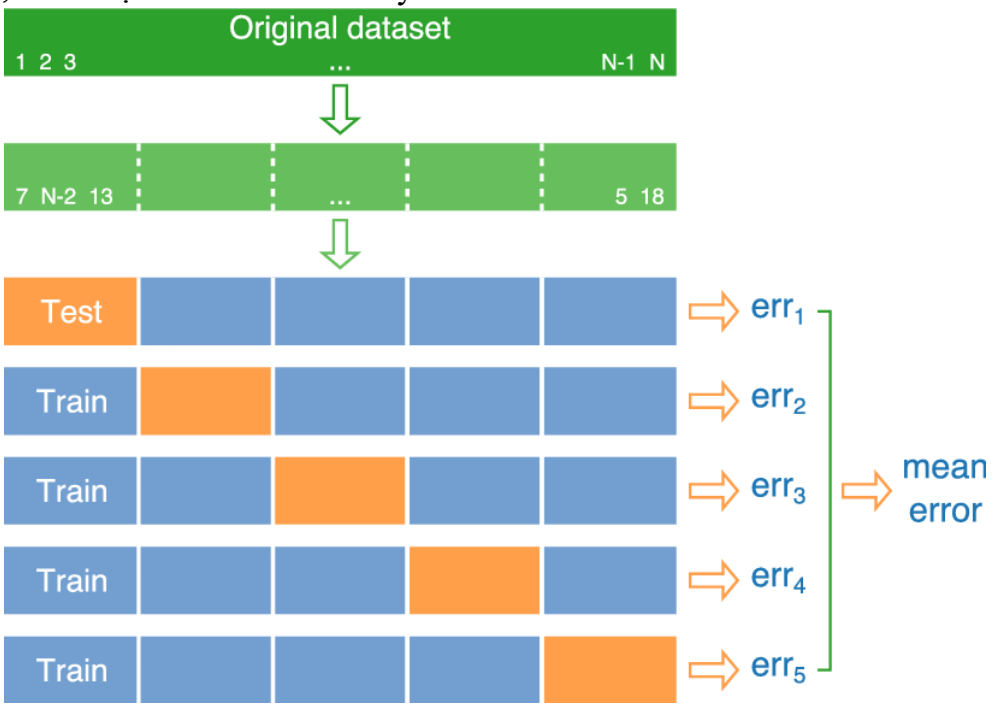
II.1. Dùng *fit()* để tìm ra các hệ số [2].

- *Dự đoán:* Sau khi đã tìm ra được hệ số và xây dựng công thức, tiến hành sử dụng các giá trị của 11 đặc trưng trong *test* để dự đoán lương, sử dụng phương thức *predict()* của thư viện để thực hiện, gán giá trị dự đoán là *y_predict1a*.

- *Tạo giá trị MAE để kiểm tra:* sử dụng hàm *mean_absolute_error()* của thư viện *sklearn*, đầu vào là giá trị *y_test* là data cho sẵn và *y_predict1a* là giá trị dự đoán sau khi thực hiện các bước trên, hàm sẽ trả về giá trị MAE.

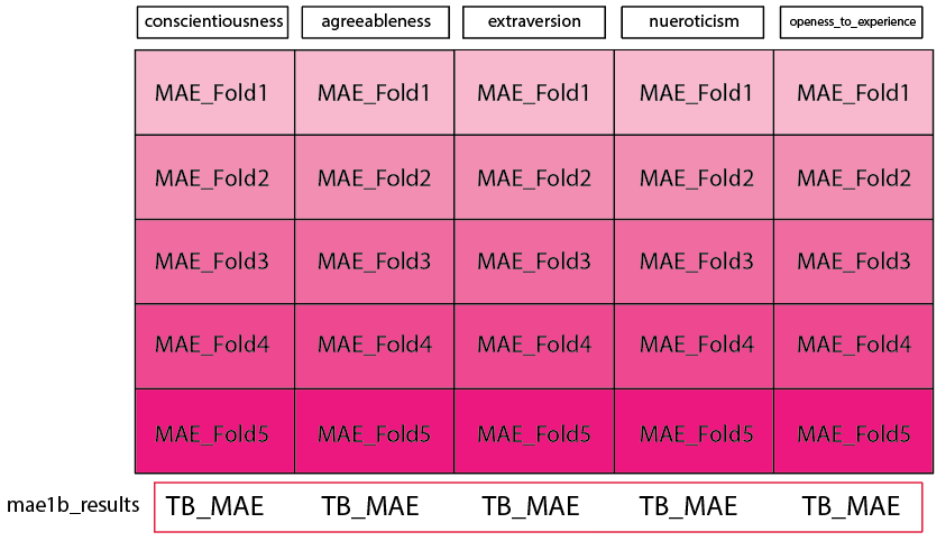
2. Yêu cầu 1b:

- *Đầu vào:* Các đặc trưng: *conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience*.
- *Mô tả hàm tìm ra hệ số:* Sử dụng *np.random.shuffle()* để xáo trộn dữ liệu đầu vào và gán vào *X_shuffle* và *Y_shuffle*, tạo list features theo tên các đặc trưng yêu cầu, sử dụng lớp *KFold()* để cắt dữ liệu theo K-Fold Cross Validation. Lớp *KFold()* của thư viện sklearn sẽ chia dữ liệu theo k phần (tham số đầu vào), mỗi phần gọi là Fold sẽ lần lượt làm tập để kiểm tra, các phần còn lại sẽ là dữ liệu để huấn luyện, mình họa như hình dưới đây:



II.2. Minh họa KFold [1]

Tạo *mae1b_results* để lưu giá trị MAE cho mỗi đặc trưng, với mỗi đặc trưng theo yêu cầu, sẽ thực hiện huấn luyện theo kiểu KFold, tức là sẽ dùng phương thức *fit()* để huấn luyện các tập huấn luyện của KFold, sau đó sẽ dùng *predict()* để dự đoán các kết quả với tập huấn luyện và tập kiểm tra, tính các MAE của từng fold, sau đó tính giá trị trung bình của từng MAE và lưu vào *mae1b_results*, minh họa như hình dưới đây:



II.3. Minh họa tính MAE cho từng đặc trưng

Sau đó sẽ huấn luyện lại và tìm ra MAE của đặc trưng tốt nhất (đặc trưng có TB_MAE trong hình trên là nhỏ nhất).

3. Yêu cầu 1c:

- *Đầu vào:* Các đặc trưng: *English, Logical, Quant*.
- *Mô tả hàm:* Tương tự yêu cầu 1b.

4. Yêu cầu 1d:

- *Mô hình 1: Mô hình 2 đặc trưng 1b 1c,* chọn 2 đặc trưng tốt nhất và dùng K-Fold để tìm ra MAE.
- *Mô hình 2: Mô hình tất cả các đặc trưng.*
- *Mô hình 3: Mô hình thêm cột tổng 3 features tốt nhất,* sẽ tìm ra 3 đặc trưng tốt nhất tương tự như câu 1b, trên tổng tất cả đặc trưng, sau đó sẽ tạo dataFrame mới là tất cả các đặc trưng cũ cộng thêm tổng 3 đặc trưng tốt nhất để tăng tính ảnh hưởng lên bộ dữ liệu.

III. Báo cáo kết quả:

1. Yêu cầu 1a:

- *Công thức hồi quy tuyến tính:*

$$Salary = -23183.330 * Gender + 702.767 * 10percentage + 1259.019 * 12percentage - 99570.608 * CollegeTier + 18369.962 * Degree + 1297.532 * collegeGPA - 8836.727 * CollegeCityTier + 141.760 * English + 145.742 * Logical + 114.643 * Quant + 34955.750 * Domain + 49248.090$$

- *MAE trên tập kiểm tra:* 105052.530

2. Yêu cầu 1b:

- *Đặc trưng tốt nhất: neuroticism*

- *Công thức hồi quy tuyến tính cho đặc trưng tốt nhất:*

$$Salary = -16021.494 * neuroticism + 304647.553$$

- *MAE các đặc trưng tương ứng mô hình k-fold Cross Validation:*

STT	Mô hình với 1 đặc trưng	MAE
1	conscientiousness	124227.061
2	agreeableness	123638.175
3	extraversion	123866.541
4	neuroticism	123594.293
5	Openness_to_experience	123884.813

- *MAE trên tập kiểm tra:* 119361.917

3. Yêu cầu 1c:

- *Đặc trưng tốt nhất:Quant*
- *Công thức hồi quy tuyến tính cho đặc trưng tốt nhất:*

$$Salary = 368.852 * Quant + 117759.729$$

- *MAE các đặc trưng tương ứng mô hình k-fold Cross Validation:*

STT	Mô hình với 1 đặc trưng	MAE
1	English	120652.839
2	Logical	119862.380
3	Quant	117166.641

- *MAE trên tập kiểm tra:* 108814.060

4. Yêu cầu 1d:

- *Đặc trưng tốt nhất: Mô hình thêm cột tổng 3 features tốt nhất*
- *MAE các đặc trưng tương ứng mô hình k-fold Cross Validation:*

STT	Mô hình	MAE
1	Mô hình 2 features câu 1b và 1c	117583.751
2	Mô hình tất cả features	111502.888
3	Mô hình thêm cột tổng 3 features tốt nhất	111502.888

- *MAE trên tập kiểm tra:* 102414.458

III. Nhận xét kết quả:

1. Yêu cầu 1b:

- *Nhận xét:* Mặc dù mô hình cho đặc trưng *neuroticism* là tốt nhất tức MAE thấp nhất, các giá trị MAE của 5 mô hình này đều gần giá trị 123 000 nên ta không thể chỉ dựa vào *neuroticism* để dự đoán cho mức lương, MAE cuối cùng so với data test thì cũng ở 119000 khá là cao so với các câu 1c và 1d.

2. Yêu cầu 1c:

- *Nhận xét:* Mô hình cho đặc trưng *Quant* là tốt nhất tức MAE thấp nhất, các giá trị MAE của 3 mô hình này có khoảng cách xa hơn so với yêu cầu 1b, MAE cuối cùng so với data test thì cũng ở 108000, có thể rằng đặc trưng này có mối quan hệ

mạnh mẽ trong việc đặt ra mức lương, hoặc có thể đặc trưng này đạt được sự kiểm chứng thống kê tốt hơn các đặc trưng khác.

3. Yêu cầu 1d:

- *Nhận xét:* Mặc dù mô hình thêm tổng 3 đặc trưng tốt nhất vào data là tốt nhất, tuy nhiên mô hình tất cả các đặc trưng cũng có mức độ MAE tương đương, nên ta thấy rằng việc dự đoán mức lương sẽ tốt nhất nếu ta kết hợp tất cả các đặc trưng, hoặc cũng có thể thêm những đặc trưng tốt nhất để tăng mức độ ảnh hưởng của chúng và dự đoán kết quả gần nhất.

IV. Tài liệu tham khảo:

[1]. *Cross-validation: evaluating estimator performance (22.08.2023) tại:* https://scikit-learn.org/stable/modules/cross_validation.html

[2]. *Thực hiện Linear Regression với Scikit-learn (22.08.2023) tại:* https://topdev.vn/blog/thuc-hien-linear-regression-voi-scikit-learn/