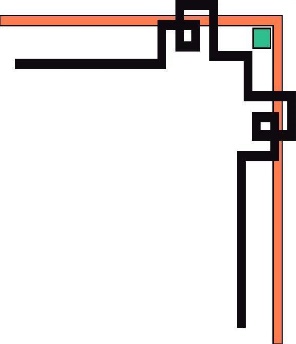
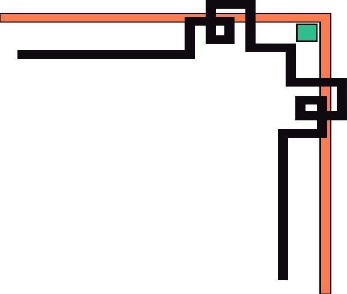
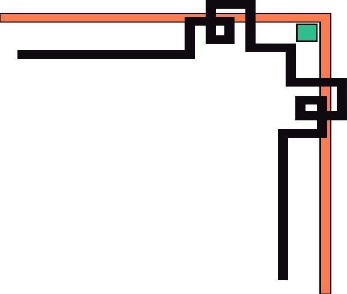
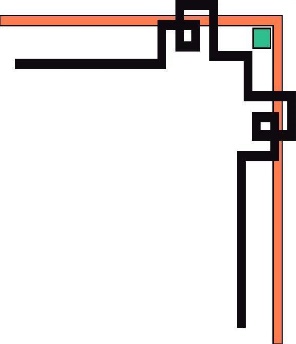
ĐẠI HỌC HUẾ



**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

A logo with text on it

Description automatically generated

**BÁO CÁO KẾT QUẢ**

**HỌC KÌ I, NĂM HỌC 2022-2023**

**HỌC PHẦN:**

**HỌC MÁY 1**

**Thừa Thiên Huế, ngày 17 tháng 12 năm 2023**

**Giảng viên hướng dẫn**: TS. Hoàng Hữu Trung

**Lớp**: Khoa học dữ liệu và Trí tuệ nhân tạo K3

**Nhóm sinh viên thực hiện**: Nhóm 5

[Bài 1: Linear Regression 2](#_Toc153736376)

[Phần 1: Giới thiệu dữ liệu 2](#_Toc153736377)

[Phần 2: Thực hiện bài toán Linear Regression 3](#_Toc153736378)

[Bài 2: Logistic Regression 12](#_Toc153736379)

[Phần 1: Giới thiệu 12](#_Toc153736380)

[1.1 Tập dữ liệu 12](#_Toc153736381)

[1.2 Thành phần dữ liệu 12](#_Toc153736382)

[1.3 Ảnh về tập dữ liệu 13](#_Toc153736383)

[PHẦN 2: Chuẩn bị dữ liệu 14](#_Toc153736384)

[PHẦN 3: Xử lí dữ liệu 16](#_Toc153736385)

[2.1 Loại bỏ dữ liệu 16](#_Toc153736386)

[2.2 Loại bỏ các giá trị ngoại lệ 17](#_Toc153736387)

[2.3 Chuẩn hóa dữ liệu 18](#_Toc153736388)

[2.4 Xử lý dữ liệu phân loại 20](#_Toc153736389)

[PHẦN 4: Thực hiện bài toán Logistic regression 21](#_Toc153736390)

[3.1 Chuẩn bị dữ liệu 21](#_Toc153736391)

[3.2 Khởi tạo và huấn luyện mô hình Logistic Regression 21](#_Toc153736392)

[3.3 Đánh giá mô hình 21](#_Toc153736393)

# Bài 1: Linear Regression

## Phần 1: Giới thiệu dữ liệu

Chúng tôi đã chọn bộ dữ liệu có tên là "Salary Dataset" để thực hiện một phân tích về mối quan hệ giữa năm kinh nghiệm và mức lương. Điều này nhằm mục đích hiểu rõ hơn về cách mức lương biến đổi theo thời gian làm việc và đồng thời có thể đưa ra các nhận định hoặc xu hướng quan trọng.

Chúng tôi chọn Google Colab để phân tích bộ dữ liệu "Salary Dataset". Với môi trường làm việc trực tuyến thuận tiện, tích hợp Python và Jupyter, khả năng xử lý mạnh mẽ, và miễn phí, Colab giúp chúng tôi tiếp cận, xử lý dữ liệu một cách hiệu quả và tiết kiệm tài nguyên.

Nguồn: [Salary\_datasets](https://www.kaggle.com/datasets/abhishek14398/salary-dataset-simple-linear-regression)

Link mã nguồn Python: [Python\_Linear regression](https://colab.research.google.com/drive/1-jfvjYnSciHjb69YztR3qg9ZefEQFnLu?hl=vi&fbclid=IwAR0diI9sJ6T3ujQOIsBu_mrzH6D-WAb_k3NwusnL88Q4aYvsfYxFQJL_TQE)

## Phần 2: Thực hiện bài toán Linear Regression

Đầu tiên, tải lên dữ liệu từ máy tính cá nhân lên máy chủ Colab.

A screenshot of a computer program

Description automatically generated

Import các thư viện cần thiết cho việc thực hiện phân tích và mô hình hóa dữ liệu bằng Linear Regression.

A screenshot of a computer program

Description automatically generated

Sử dụng *df.info()* là một bước quan trọng để đạt được cái nhìn tổng quan về dữ liệu trong DataFrame bao gồm: Tổng số dòng và cột, tên và kiểu dữ liệu của mỗi cột, số lượng giá trị không rỗng, tổng số bộ nhớ chiếm dụng.

A screenshot of a computer

Description automatically generated

Sử dụng *df.head()* hiển thị một số dòng đầu tiên của DataFrame để có cái nhìn nhanh chóng về cấu trúc và nội dung của dữ liệu.

A screenshot of a computer

Description automatically generated

Sử dụng *df* in ra toàn bộ DataFrame để có cái nhìn chi tiết.

A screenshot of a computer

Description automatically generatedA screenshot of a table

Description automatically generated

Tạo một biểu đồ phân tán.



Kết quả:

A graph with blue dots

Description automatically generated

Theo như biểu đồ, những người có nhiều năm kinh nghiệm thường có mức lương cao hơn, cũng có một số người có ít năm kinh nghiệm nhưng mức lương cao.

Lọc lấy cột đầu tiên trong df (biến độc lập)

A screenshot of a computer code

Description automatically generated

Lọc lấy cột cuối cùng trong df (biến phụ thuộc)

A screenshot of a computer code

Description automatically generated

Chia tập dữ liệu thành tập huấn luyện và tập kiểm tra.



*x\_train* là biến đại diện cho tập dữ liệu huấn luyện. *x\_train* chứa giá trị của các biến độc lập được sử dụng để "đào tạo" mô hình, giúp nó học từ dữ liệu và có khả năng dự đoán tốt trên dữ liệu mới.

A screenshot of a test

Description automatically generated

*x\_test* là biến được sử dụng để đại diện cho tập dữ liệu kiểm tra. Nó chứa giá trị của các biến độc lập mà mô hình chưa bao giờ nhìn thấy trong quá trình đào tạo. *x\_test* giúp đánh giá hiệu suất của mô hình trên dữ liệu mới và kiểm tra khả năng tổng quát hóa của nó.

A screenshot of a computer

Description automatically generated

*y\_train* là biến được sử dụng để đại diện cho tập dữ liệu kết quả huấn luyện. *y\_train* chứa giá trị của biến phụ thuộc (kết quả) tương ứng với *x\_train* - tức là giá trị mà mô hình cần học cách dự đoán từ các biến độc lập.

A screenshot of a computer

Description automatically generated

*y\_test* là biến được sử dụng để đại diện cho tập dữ liệu kết quả kiểm tra. *y\_test* chứa giá trị của biến phụ thuộc (kết quả) tương ứng với **x\_test**. Điều này giúp đánh giá khả năng dự đoán của mô hình trên dữ liệu mới, chưa được mô hình nhìn thấy trong quá trình đào tạo.

A screenshot of a computer

Description automatically generated

Khởi tạo một mô hình tuyến tính và huấn luyện mô hình Linear Regression trên tập dữ liệu huấn luyện.

A screenshot of a computer code

Description automatically generated

Dự đoán giá trị của biến phụ thuộc (kết quả) trên tập dữ liệu kiểm tra *(x\_test).*

A screenshot of a computer code

Description automatically generated

*a = lr.coef\_* được sử dụng để lấy giá trị hệ số (coefficient) từ mô hình Linear Regression đã được đào tạo *(lr)*.

*b = lr.intercept\_* được sử dụng để lấy giá trị của hệ số chệch (intercept) từ mô hình Linear Regression đã được đào tạo *(lr).* Hệ số chệch là giá trị dự đoán của biến phụ thuộc khi tất cả các biến độc lập đều bằng 0.

A screenshot of a computer code

Description automatically generated

Thay số cho ra kết quả.

A screenshot of a computer

Description automatically generated

Hiển thị các thống kê mô tả của DataFrame bao gồm số lượng, mean, std, min, max và các phần vị của các biến trong DataFrame.

A screenshot of a computer

Description automatically generated

Thực hiện kiểm định Shapiro-Wilk để kiểm tra tính chuẩn của phân phối của biến 'Salary'. Mục đích của kiểm định Shapiro-Wilk trong trường hợp này là kiểm tra xem biến 'Salary' có phân phối gần với phân phối chuẩn hay không, điều này có thể ảnh hưởng đến sự hiệu quả của các phương pháp thống kê dựa trên giả định phân phối chuẩn.

A screenshot of a computer code

Description automatically generated

Tạo biểu đồ histogram của biến 'Salary' và cùng lúc vẽ đường phân phối ước lượng (Kernel Density Estimate - KDE) lên trên histogram.



Kết quả:

A graph with blue columns

Description automatically generated

Biểu đồ này cho thấy phân bố tần số của mức lương trong dữ liệu. Đường KDE cho thấy rằng mức lương có xu hướng tập trung ở khoảng 50.000 đến 100.000 đô la.

Tạo ma trận tương quan và tạo biểu đồ heatmap của ma trận tương quan đó.



Kết quả:

A red and blue squares with white text

Description automatically generated

Biểu đồ này cho thấy mối quan hệ giữa các biến trong DataFrame. Các giá trị dương cho thấy mối quan hệ thuận, trong khi các giá trị âm cho thấy mối quan hệ nghịch. Cường độ của mối quan hệ được biểu thị bằng độ đậm của màu.

Dự đoán mức lương cho năm kinh nghiệm việc làm:

A computer screen shot of a program code

Description automatically generated

Kết quả:

A graph with a line and a red arrow

Description automatically generated

Biểu đồ này được sử dụng để hiển thị mối quan hệ giữa kinh nghiệm làm việc và mức lương. Đường màu xanh dương cho thấy đường hồi quy tuyến tính giữa hai biến. Mũi tên đỏ cho thấy mức lương dự đoán (77767$) cho năm kinh nghiệm làm việc là 5,5 năm.

# Bài 2: Logistic Regression

## Phần 1: Giới thiệu

### 1.1 Tập dữ liệu

Tập dữ liệu nhóm chọn có tên **weatherAUS.csv**.

Nguồn: [Rain in Australia (kaggle.com)](https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package)

Link mã nguồn Python: [Python\_Logistic\_regression](https://colab.research.google.com/drive/1yRYRr9JLi3HMOxQg1GRNfsHEVgYMLLNJ#scrollTo=pQO-4hSO28oU)

#### 1.1.1 Ngữ cảnh

Dự đoán mưa ngày hôm sau bằng cách đào tạo các mô hình phân loại trên biến mục tiêu RainTomorrow.

#### 1.1.2 Nội dung

Bộ dữ liệu này chứa khoảng 10 năm quan sát thời tiết hàng ngày từ nhiều địa điểm trên khắp nước Úc.

RainTomorrow là biến mục tiêu để dự đoán. Nó có nghĩa là - trời mưa vào ngày hôm sau, Có hay Không? Cột này là Có nếu mưa cho ngày hôm đó là 1mm trở lên.

### 1.2 Thành phần dữ liệu

Bộ dữ liệu này chứa khoảng 10 năm quan sát thời tiết hàng ngày từ nhiều trạm thời tiết Úc.

RainTomorrow là biến mục tiêu để dự đoán. Nó có nghĩa là - trời mưa vào ngày hôm sau, Có hay Không?

Cột này là Có nếu mưa cho ngày hôm đó là 1mm trở lên.

|  |  |
| --- | --- |
| **Tên cột** | **Giải thích** |
| Date | Ngày quan sát |
| Location | Tên chung của vị trí trạm thời tiết |
| MinTemp | Nhiệt độ tối thiểu (°C) |
| MaxTemp | Nhiệt độ tối đa (°C) |
| Rainfall | Lượng mưa ghi nhận trong ngày (mm) |
| Evaporation | Sự bốc hơi của chảo loại A (mm) trong 24 giờ đến 9 giờ sáng |
| Sunshine | Số giờ nắng chói chang trong ngày (giờ) |
| WindGustDir | Hướng gió giật mạnh nhất trong 24 giờ đến nửa đêm |
| WindGustSpeed | Tốc độ (km/h) gió giật mạnh nhất trong 24 giờ đến nửa đêm |
| WindDir9am | Hướng gió lúc 9 giờ sáng |
| WindDir3pm | Hướng gió lúc 3 giờ chiều |
| WindSpeed9am | Tốc độ gió (km/giờ) trung bình trên 10 phút trước 9 giờ sáng |
| WindSpeed3pm | Tốc độ gió (km/giờ) trung bình trong 10 phút trước 3 giờ chiều |
| Humidity9am | Độ ẩm (%) lúc 9 giờ sáng |
| Humidity3pm | Độ ẩm (%) lúc 3 giờ chiều |
| Pressure9am | Áp suất khí quyển (hpa) giảm xuống mực nước biển trung bình lúc 9 giờ sáng |
| Pressure3pm | Áp suất khí quyển (hpa) giảm xuống mực nước biển trung bình lúc 3 giờ chiều |
| Cloud9am | Một phần bầu trời bị mây che khuất lúc 9 giờ sáng. Điều này được đo bằng "oktas", là đơn vị của phần tám. Nó ghi lại bao nhiêu |
| Cloud3pm | Phần bầu trời bị mây che khuất (trong "oktas": phần tám) lúc 3 giờ chiều. Xem Cload9am để biết mô tả về các giá trị |
| Temp9am | Nhiệt độ ((°C) lúc 9 giờ sáng |
| Temp3pm | Nhiệt độ ((°C) lúc 3 giờ chiều |
| RainToday | Boolean: 1 nếu lượng mưa (mm) trong 24 giờ đến 9 giờ sáng vượt quá 1mm, nếu không thì 0 |
| RainTomorrow | Lượng mưa ngày hôm sau tính bằng mm. Được sử dụng để tạo biến phản hồi RainTomorrow. Một loại thước đo "rủi ro". |

### 1.3 Ảnh về tập dữ liệu

A screen shot of a white sheet

Description automatically generated

## PHẦN 2: Chuẩn bị dữ liệu

Nhập thư viện để chuẩn bị cho các bước xử lý dữ liệu sau này.

A screenshot of a computer

Description automatically generated

Dùng hàm warnings để tắt tất cả các cảnh báo trong quá trình chạy mã.

A white rectangular object with red green and black text

Description automatically generated

Đọc file CSV và lưu DataFrame vào biến **df**. Sau đó in ra màn hình kích thước của DataFrame.

A screenshot of a computer

Description automatically generated

Kết quả:

A screenshot of a computer

Description automatically generated

In ra màn hình tên các cột.

A white rectangular object with black text

Description automatically generated

Kết quả:

A white background with black text

Description automatically generated

In ra các thống kê cơ bản như mean, std, min, max của mỗi cột dạng số trong DataFrame.

A white rectangular object with green text

Description automatically generated

Kết quả:

A screenshot of a computer

Description automatically generated

## PHẦN 3: Xử lí dữ liệu

### 2.1 Loại bỏ dữ liệu

In ra số lượng giá trị không rỗng (không phải là null) của từng cột, sắp xếp theo giá trị tăng dần.

A white box with green text

Description automatically generated

Kết quả:

A screenshot of a computer

Description automatically generated

Loại bỏ các cột không cần thiết từ DataFrame. Các cột bị loại bỏ là 'Sunshine', 'Evaporation', 'Cloud3pm', 'Cloud9am', và 'Date' và in ra kích thước mới của DataFrame.

A screen shot of a computer

Description automatically generated

Loại bỏ các hàng có ít nhất một giá trị null từ DataFrame và in ra kích thước mới của DataFrame sau khi loại bỏ giá trị null.

A white rectangular object with green text

Description automatically generated

### 2.2 Loại bỏ các giá trị ngoại lệ

Sử dụng Z-score để xác định giá trị ngoại lệ và loại bỏ chúng khỏi DataFrame. In ra kích thước mới của DataFrame sau khi loại bỏ giá trị ngoại lệ.

A screen shot of a computer code

Description automatically generated

Kết quả:

A screenshot of a computer screen

Description automatically generated

### 2.3 Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu số bằng cách sử dụng hàm scale từ thư viện preprocessing của scikit-learn.

A screen shot of a computer code

Description automatically generated

Kết quả:

A screenshot of a computer

Description automatically generated

Sử dụng LabelEncoder để chuyển các giá trị "Yes" và "No" trong cột "RainToday" và "RainTomorrow" thành 1 và 0 tương ứng.

A screen shot of a computer code

Description automatically generated

Kết quả:

A screenshot of a computer

Description automatically generated

### 2.4 Xử lý dữ liệu phân loại

In ra số lượng và tên của các biến phân loại trong DataFrame.

A screen shot of a computer code

Description automatically generated

Kết quả:



In ra các giá trị duy nhất trong từng cột phân loại và tạo các biến giả tưởng cho các biến phân loại sử dụng phương pháp "one-hot encoding" bằng hàm get\_dummies của pandas.

A screenshot of a computer code

Description automatically generated

Kết quả:

A screenshot of a computer

Description automatically generated

## PHẦN 4: Thực hiện bài toán Logistic regression

### 3.1 Chuẩn bị dữ liệu

#### 3.1.1 Tính năng mô hình

Đặt X là tất cả các đặc trưng và y là biến mục tiêu:

* X là DataFrame chứa tất cả các đặc trưng (các cột trừ cột "RainTomorrow");
* y là Series chứa biến mục tiêu "RainTomorrow".

A screen shot of a computer

Description automatically generated

#### 3.1.2 Chia dữ liệu thành tập huấn luyện và tập kiểm thử

Sử dụng hàm train\_test\_split từ scikit-learn để chia dữ liệu thành tập huấn luyện và tập kiểm thử.

A screenshot of a computer

Description automatically generated

### 3.2 Khởi tạo và huấn luyện mô hình Logistic Regression

Sử dụng mô hình Logistic Regression và huấn luyện nó trên tập huấn luyện (X\_train, y\_train) và y\_pred.

A screenshot of a computer

Description automatically generated

### 3.3 Đánh giá mô hình

Dùng mô hình đã huấn luyện để dự đoán trên tập kiểm thử và tính độ chính xác của mô hình.

A computer screen shot of a computer code

Description automatically generated

Kết quả:



Chọn ngẫu nhiên 100 điểm từ tập kiểm thử để biểu diễn trên biểu đồ.

A white rectangular object with green text

Description automatically generated

Vẽ biểu đồ scatter plot so sánh giữa kết quả thực tế và kết quả dự đoán.

A screenshot of a computer code

Description automatically generated

Kết quả:

A white rectangular frame with red and black dots

Description automatically generated