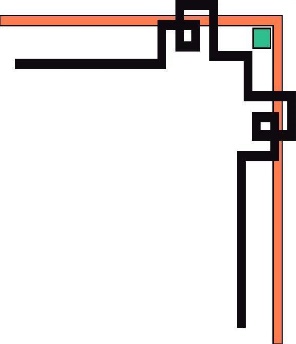
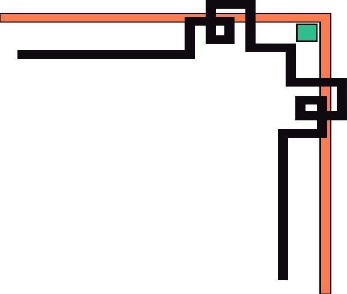
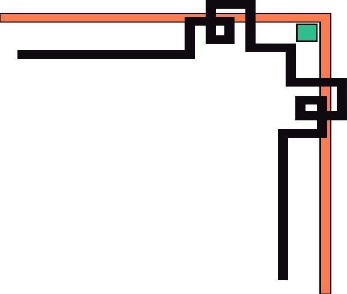
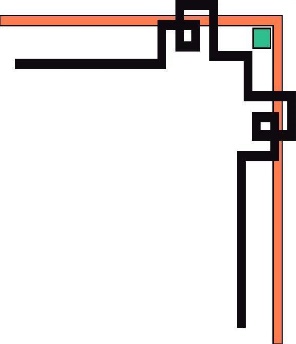
ĐẠI HỌC HUẾ



**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 13 tháng 12 năm 202**

**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

# MỞ ĐẦU

## GIỚI THIỆU DỮ LIỆU

### 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:

### 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

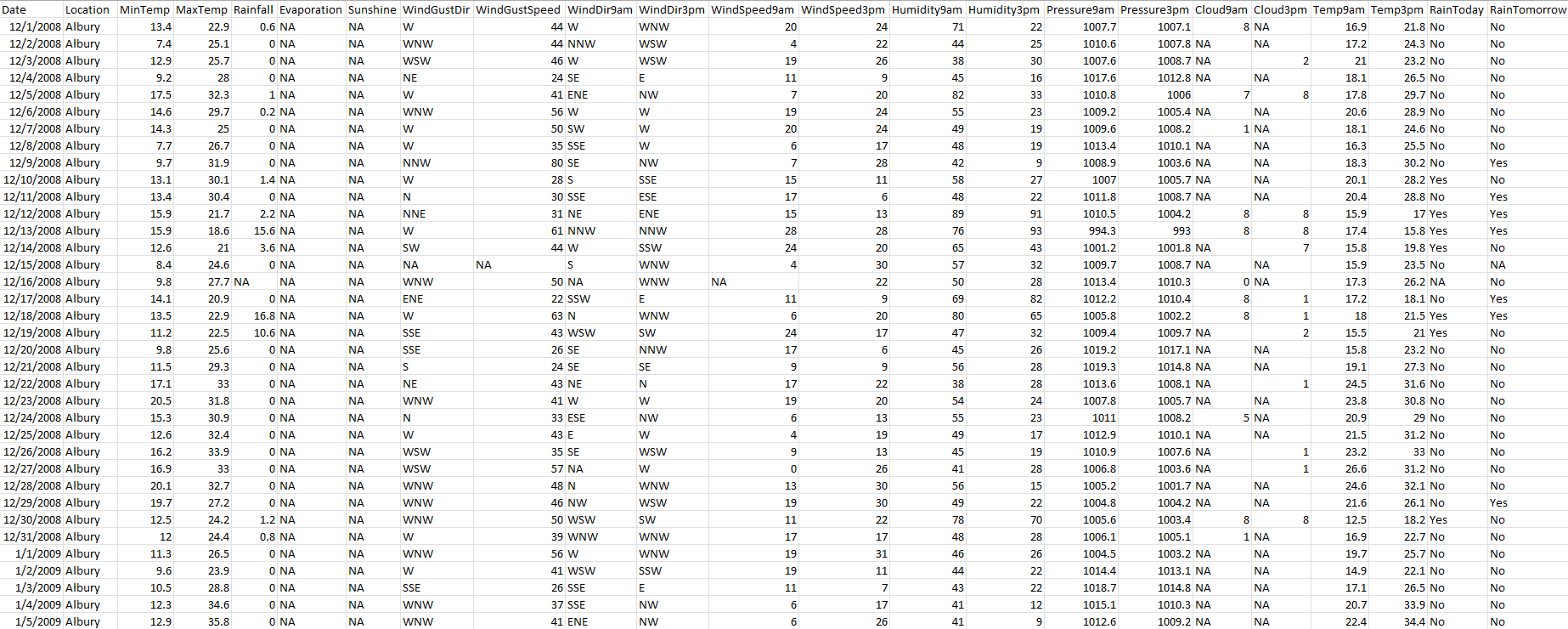
Bộ dữ liệu này chứa khoảng 10 năm quan sát thời tiết hàng ngày từ nhiều trạm thời tiết Úc.

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



# TIẾN TRÌNH

## PHẦN 1: 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:

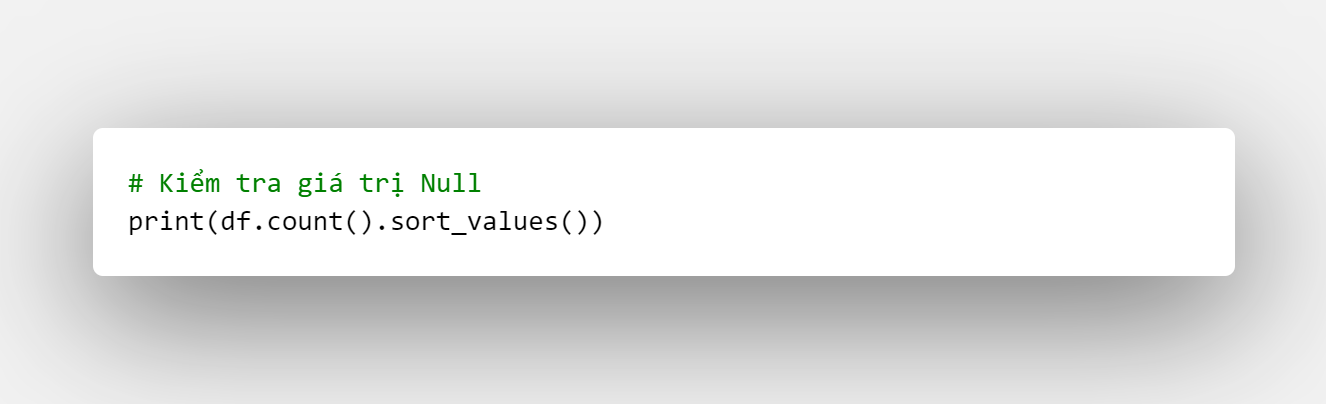
A screenshot of a computer

Description automatically generated

## PHẦN 2: XỬ LÝ 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.



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 3: 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