| **ĐẠI HỌC BÁCH KHOA HÀ NỘI  TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**  **---□**&**□---**      **BÁO CÁO BÀI TẬP LỚN**  **Nhập Môn AI**  **Đề tài: Dự báo thời tiết**  **GVHD: TS. Đỗ Tiến Dũng**  **Nhóm 8 - 154016 - IT3160**  Hoả Đức Việt 20205046  Lê Duy Quý 20205018  Nhuien Tkhi Kam Tu 20210988  Nguyễn Việt Anh 20215307  **Hà Nội, tháng 12 năm 2024** |
| --- |

[**1. Mô tả bài toán 3**](#_no5of9a6vl0t)

[**2. Thu thập dữ liệu 3**](#_w089i2v5j8d1)

[**3. Tiền xử lý dữ liệu 5**](#_b1x7rfkyu55l)

[- Các biểu đồ đánh giá tổng quát 7](#_45bocst7n8ln)

[- Biểu đồ phân tích đặc trưng 7](#_w4sb468he1jq)

[Kết luận 10](#_tal7bmasia80)

[Tổng Quan Dữ Liệu 10](#_2iy2p2t3bms7)

[Chất Lượng Dữ Liệu 10](#_fj5i90aga6r4)

[**4. Xây dựng mô hình và huấn luyện 11**](#_19xrt6kz1ly)

[4.1. Đọc dữ liệu và chọn các trường dùng để huấn luyện mô hình 11](#_fuduj9xb60su)

[4.2 Chia dữ liệu 11](#_jq9orhj5vouq)

[4.3 Lập trình mô hình 12](#_sxgue9tblv4d)

[4.4 Huấn luyện và kiểm tra mô hình 14](#_hdg6qxuxuwmn)

[**5. Đánh giá mô hình 15**](#_qz3vzgcd7267)

[5.1. Đánh giá về thuật toán được lựa chọn K-NN 15](#_j2d1xfm60qos)

[a. Ưu điểm của thuật toán K-NN trong dự án 15](#_untd62tk340)

[b. Nhược điểm của thuật toán K-NN 15](#_f3p09yl8k5vp)

[5.2. Đánh giá về sự thay đổi của k và độ chính xác (accuracy) 15](#_smlf1xmur0zy)

[5.3.Các biểu đồ đánh giá mô hình (với k =55) 17](#_5hfx6xe8izwf)

[**6. Kết luận 20**](#_gmndpktwvnba)

# 

| Họ tên | MSSV | Phân công công việc | Tỷ trọng |
| --- | --- | --- | --- |
| Hoả Đức Việt | 20205046 | * Tiền xử lý dữ liệu, báo cáo, code mô hình | 25% |
| Lê Duy Quý | 20205018 | * Code mô hình, báo cáo | 25% |
| Nhuien Tkhi Kam Tu | 20210988 | * Thu thập dữ liệu, báo cáo, slide | 25% |
| Nguyễn Việt Anh | 20215307 | * Đánh giá mô hình, báo cáo | 25% |

Link mã nguồn: [*AI\_Project\_Weather\_Forecast*](https://colab.research.google.com/drive/18KGL4H5A5xEG2mXaJ055sVc8i-cMI41r?usp=sharing)

Tập dữ liệu:[*Data\_ AI\_Project\_Weather\_Forecast*](https://drive.google.com/file/d/1CXpN329l5xxHjN33R6T9itMmUphE131z/view?usp=drive_link)

# **1. Mô tả bài toán**

Dự báo mưa là một phần quan trọng trong lĩnh vực khí tượng, giúp cung cấp thông tin cần thiết cho các hoạt động hàng ngày và phòng chống thiên tai. Đề tài này tập trung vào việc phát triển một mô hình dự báo mưa sử dụng thuật toán K-Nearest Neighbors (KNN), dựa trên các yếu tố khí hậu như nhiệt độ, độ ẩm, tốc độ gió, và mức độ che phủ mây. Mô hình KNN được lựa chọn do tính đơn giản và hiệu quả trong dự báo, giúp xác định khả năng mưa bằng cách tìm kiếm các điều kiện thời tiết tương tự trong quá khứ. Bộ dữ liệu được sử dụng để huấn luyện và kiểm thử mô hình bao gồm thông tin thời tiết lịch sử chi tiết từ nhiều năm, thu thập từ các nguồn dữ liệu trực tuyến đáng tin cậy và cơ sở dữ liệu khí tượng. Kết quả của dự án này không chỉ cung cấp cái nhìn sâu sắc về mối quan hệ giữa các yếu tố thời tiết và lượng mưa, mà còn gợi mở tiềm năng áp dụng các phương pháp dự báo đơn giản nhưng mạnh mẽ trong việc cải thiện dự báo thời tiết và ứng phó với các hiện tượng thời tiết cực đoan trong tương lai.

# **2. Thu thập dữ liệu**

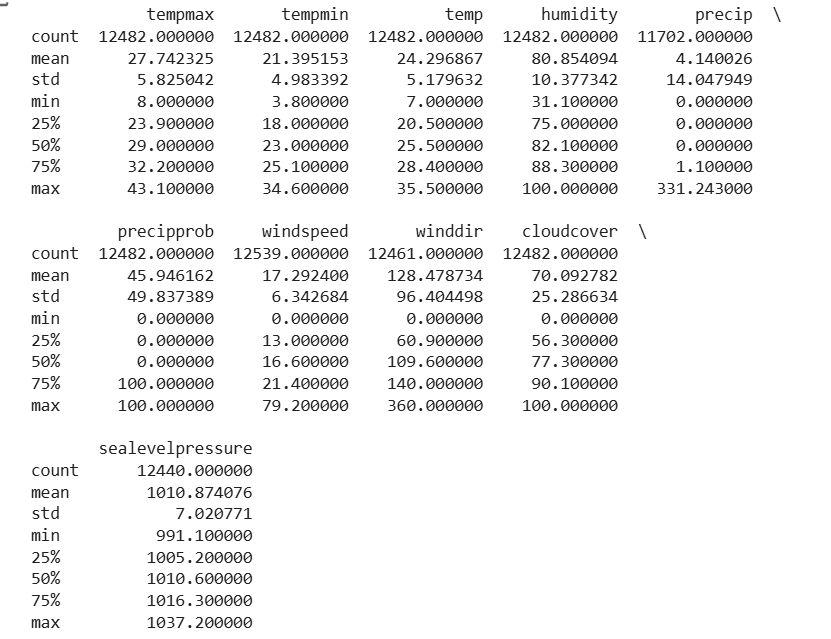
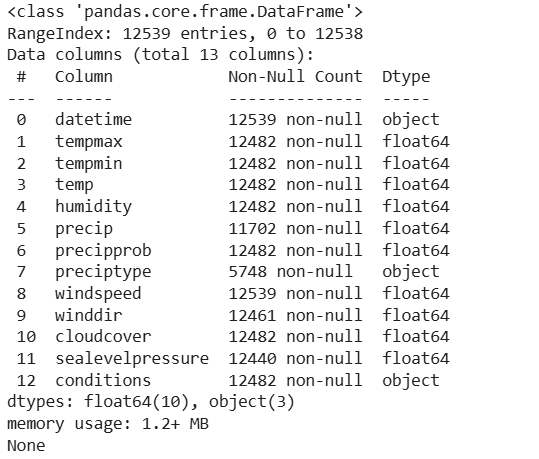
* Xác định mục tiêu, nhu cầu thu thập dữ liệu
* Xác định các nguồn dữ liệu
* Lên kế hoạch thu thập dữ liệu
* Cào dữ liệu
* Lưu trữ dữ liệu
* <https://www.kaggle.com/>

(*link lấy data:* [*https://www.kaggle.com/datasets/khinguyntin/hanoiweather1990-2024-data-from-visual-crossing*](https://www.kaggle.com/datasets/khinguyntin/hanoiweather1990-2024-data-from-visual-crossing)

Thông tin Dữ liệu:

Bộ dữ liệu có 12,539 dòng và 13 cột, bao gồm các thuộc tính liên quan đến thời tiết của Hà Nội qua nhiều năm.

Các cột chứa thông tin về ngày (datetime), nhiệt độ cao nhất (tempmax), nhiệt độ thấp nhất (tempmin), nhiệt độ trung bình (temp), độ ẩm (humidity), lượng mưa (precip), xác suất mưa (precipprob), loại mưa (preciptype), tốc độ gió (windspeed), hướng gió (winddir), độ che phủ mây (cloudcover), áp suất khí quyển ở mực nước biển (sealevelpressure), và điều kiện thời tiết (conditions).

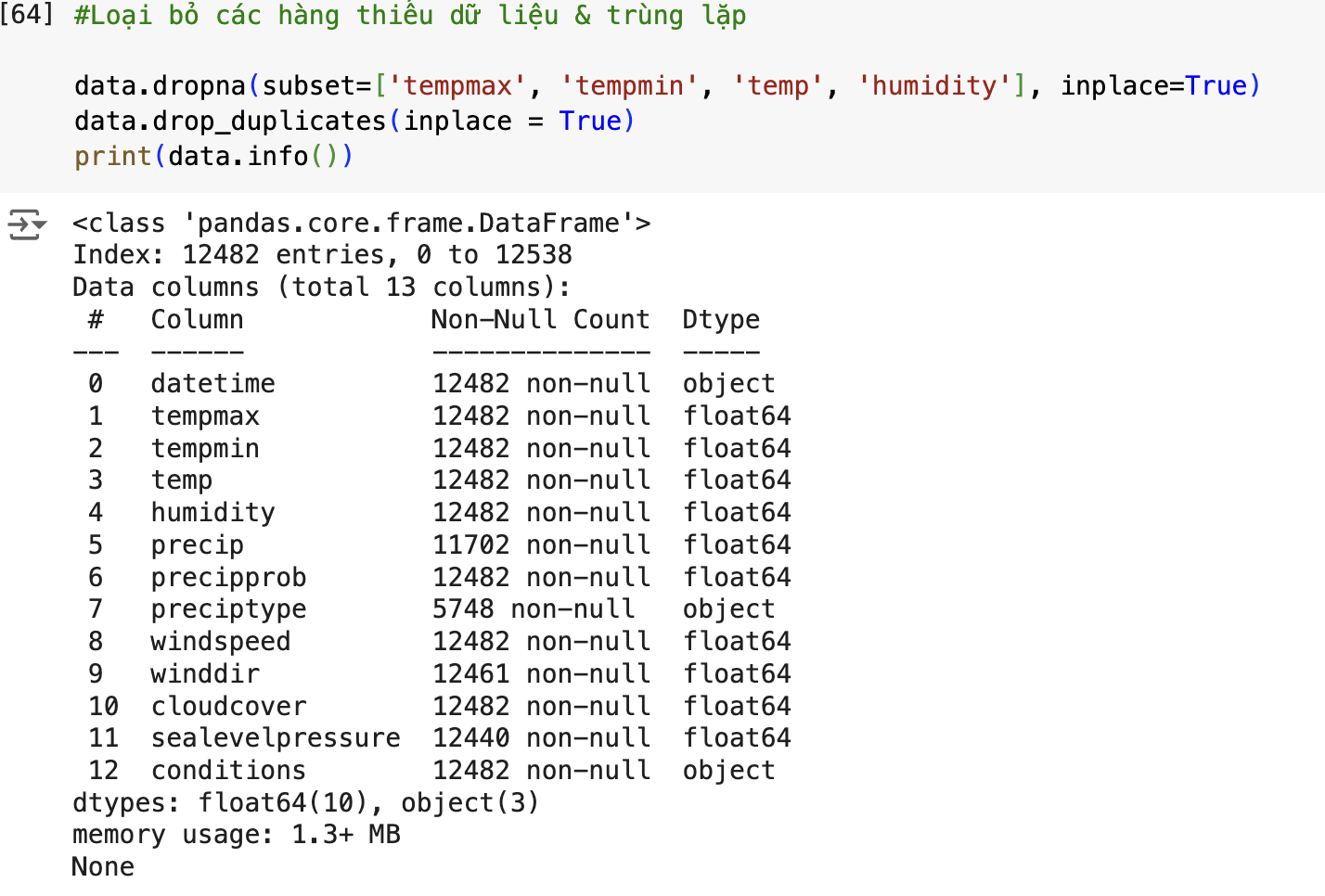


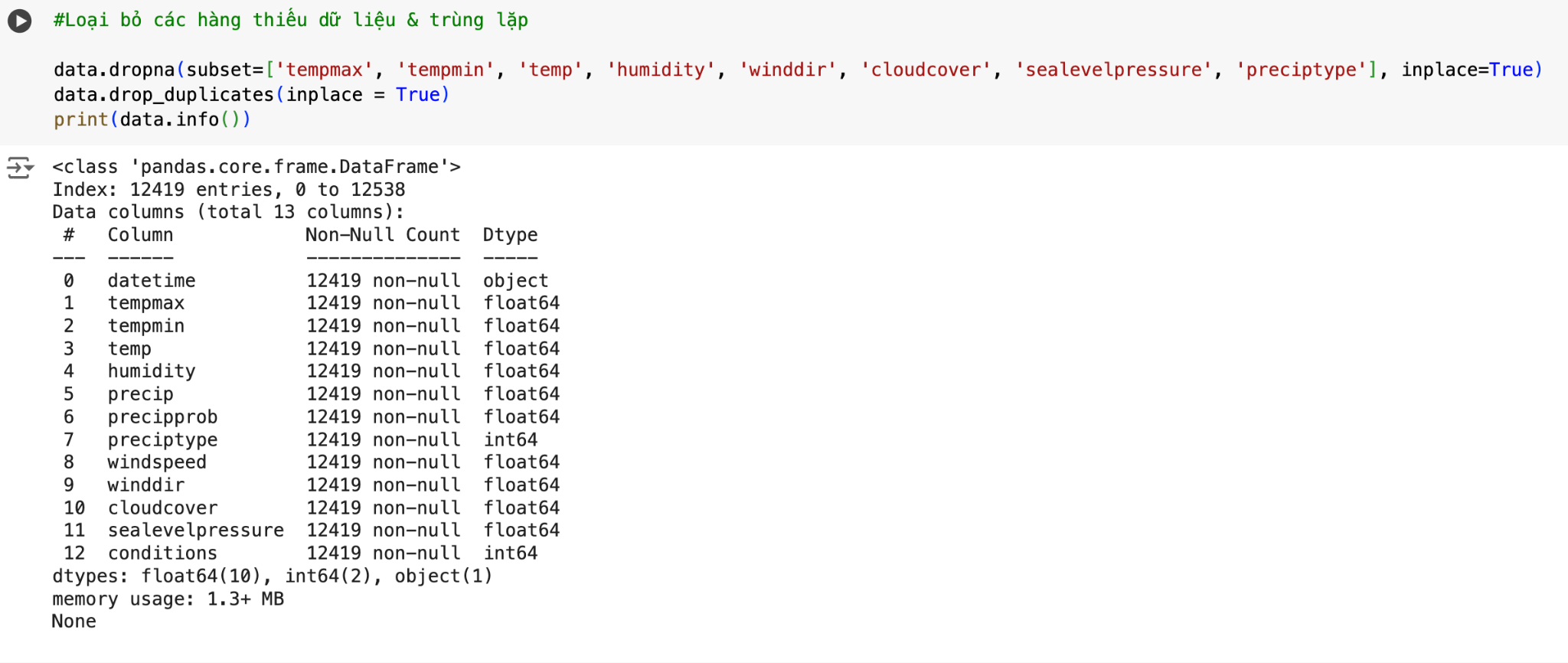
# 

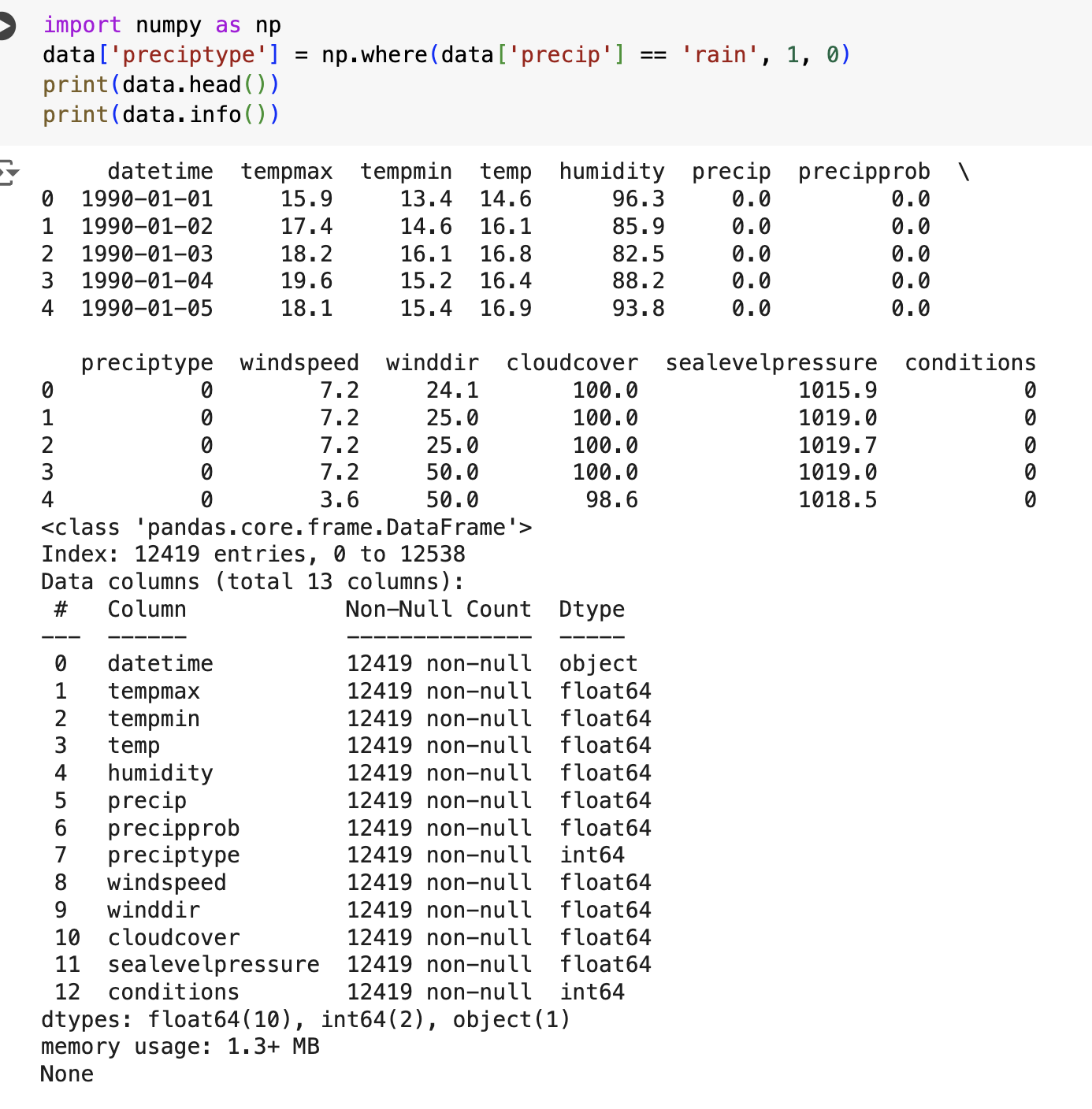
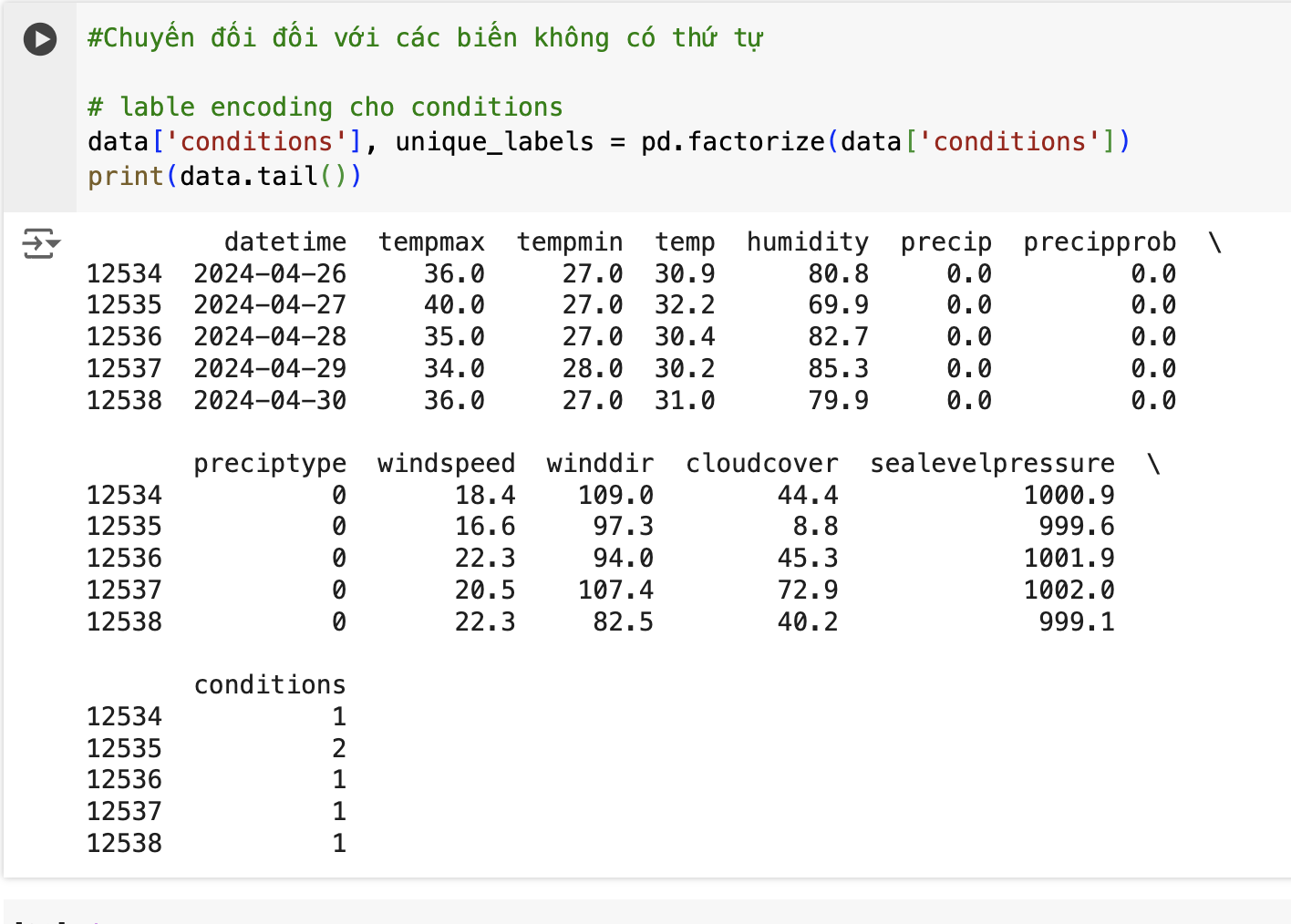
# **3. Tiền xử lý dữ liệu**

<https://colab.research.google.com/drive/18KGL4H5A5xEG2mXaJ055sVc8i-cMI41r#scrollTo=aLiS2vZcks9Z>

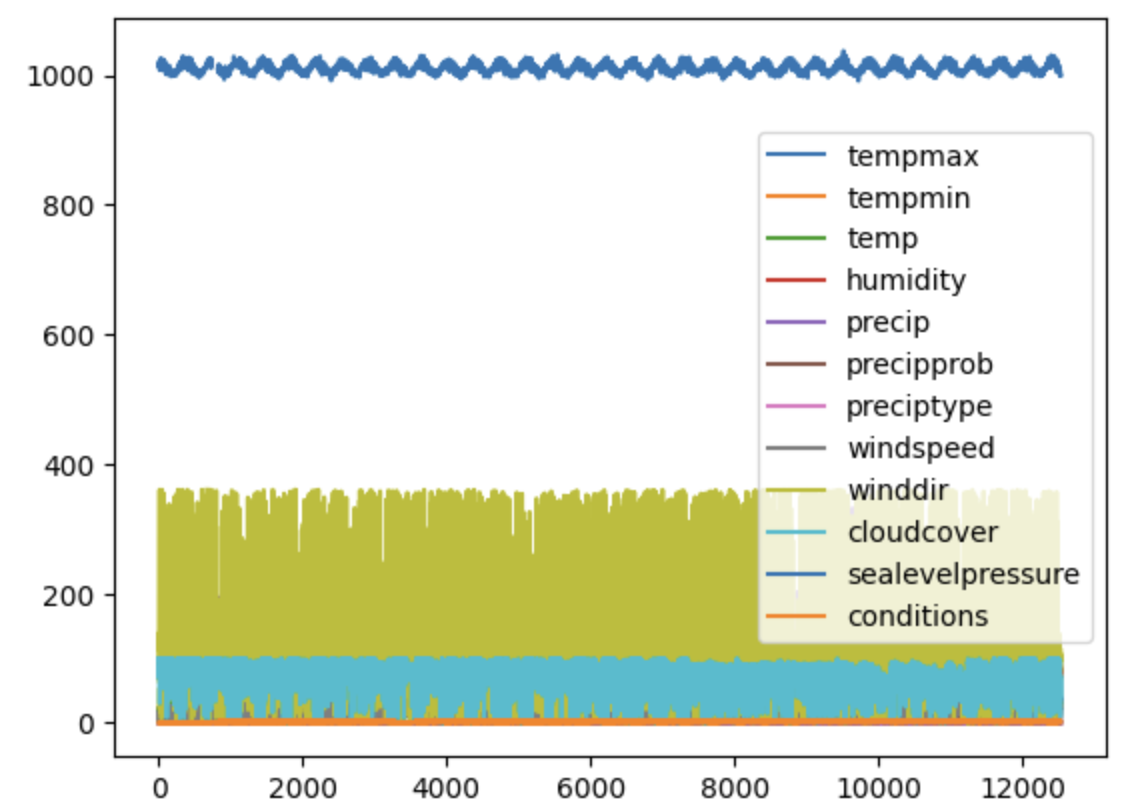
* Làm đẹp dữ liệu



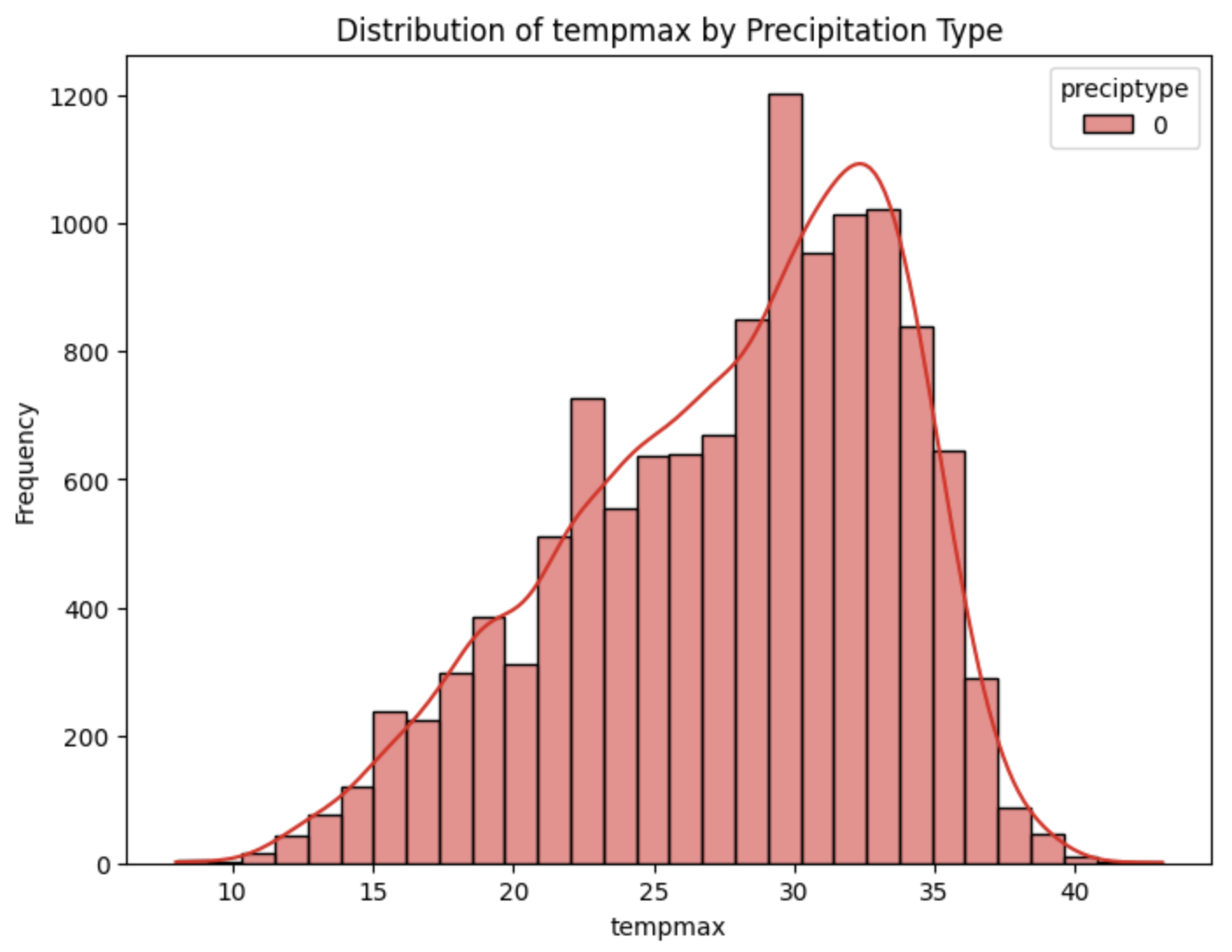
* Chuẩn hóa dữ liệu 

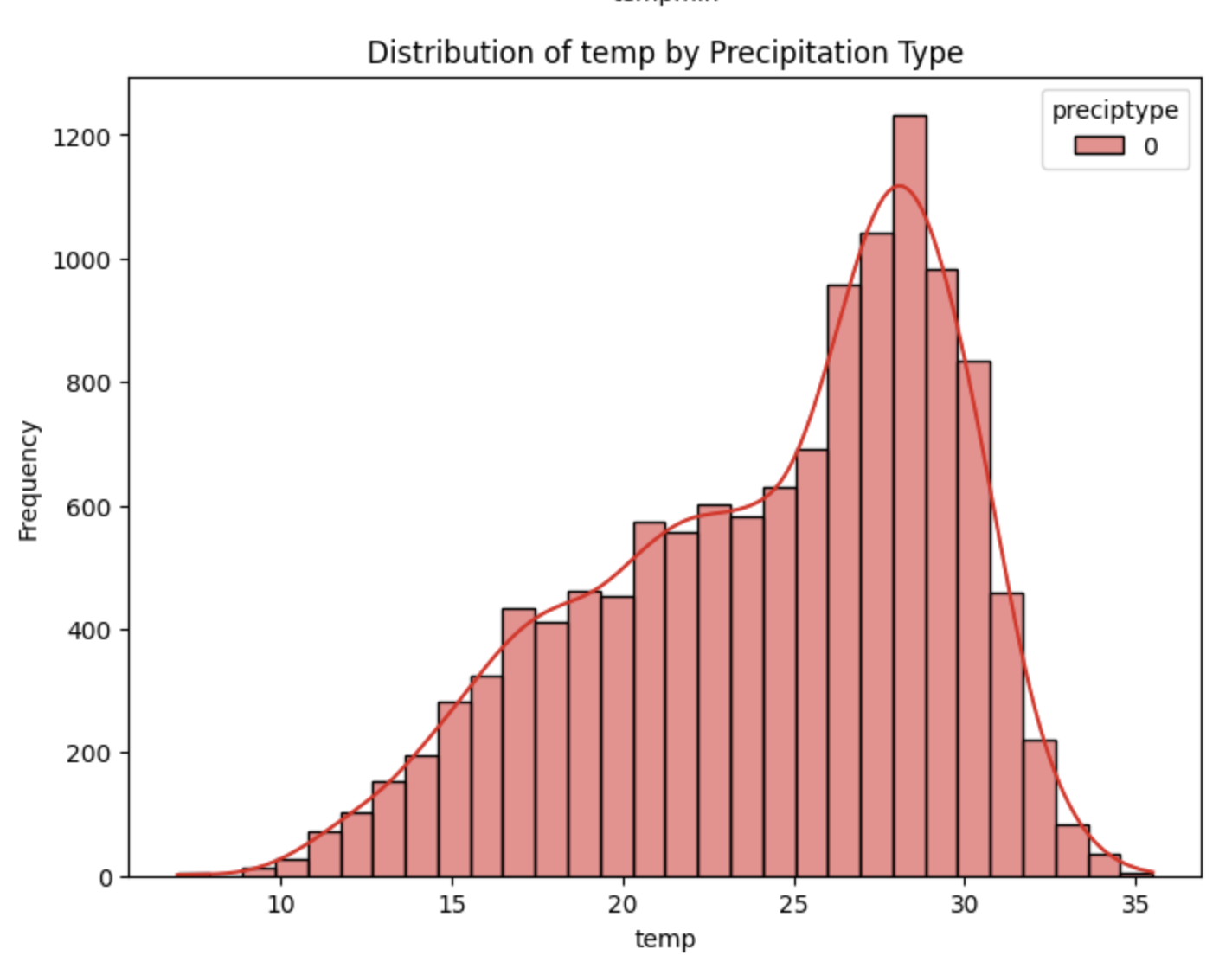


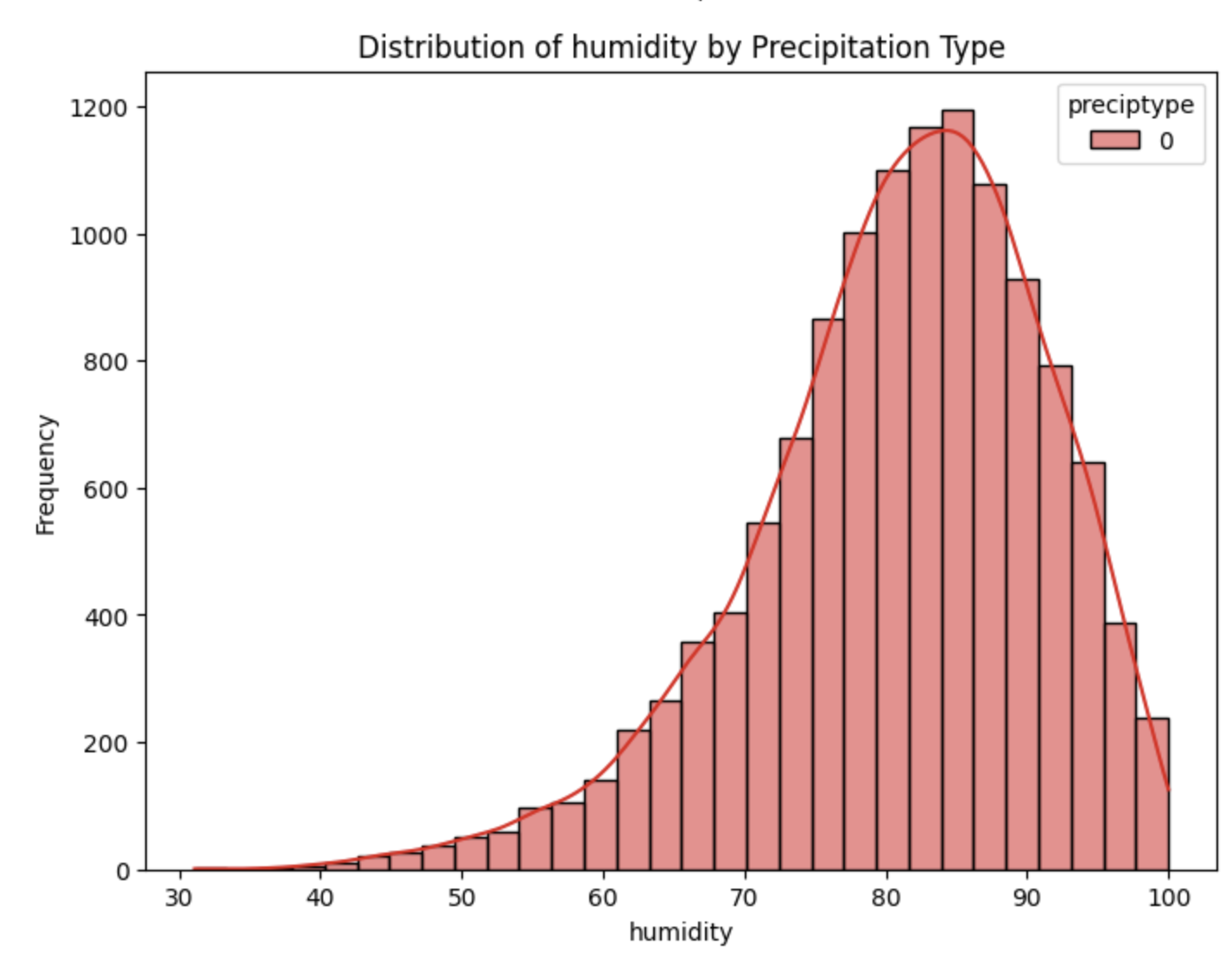
## Các biểu đồ đánh giá tổng quát

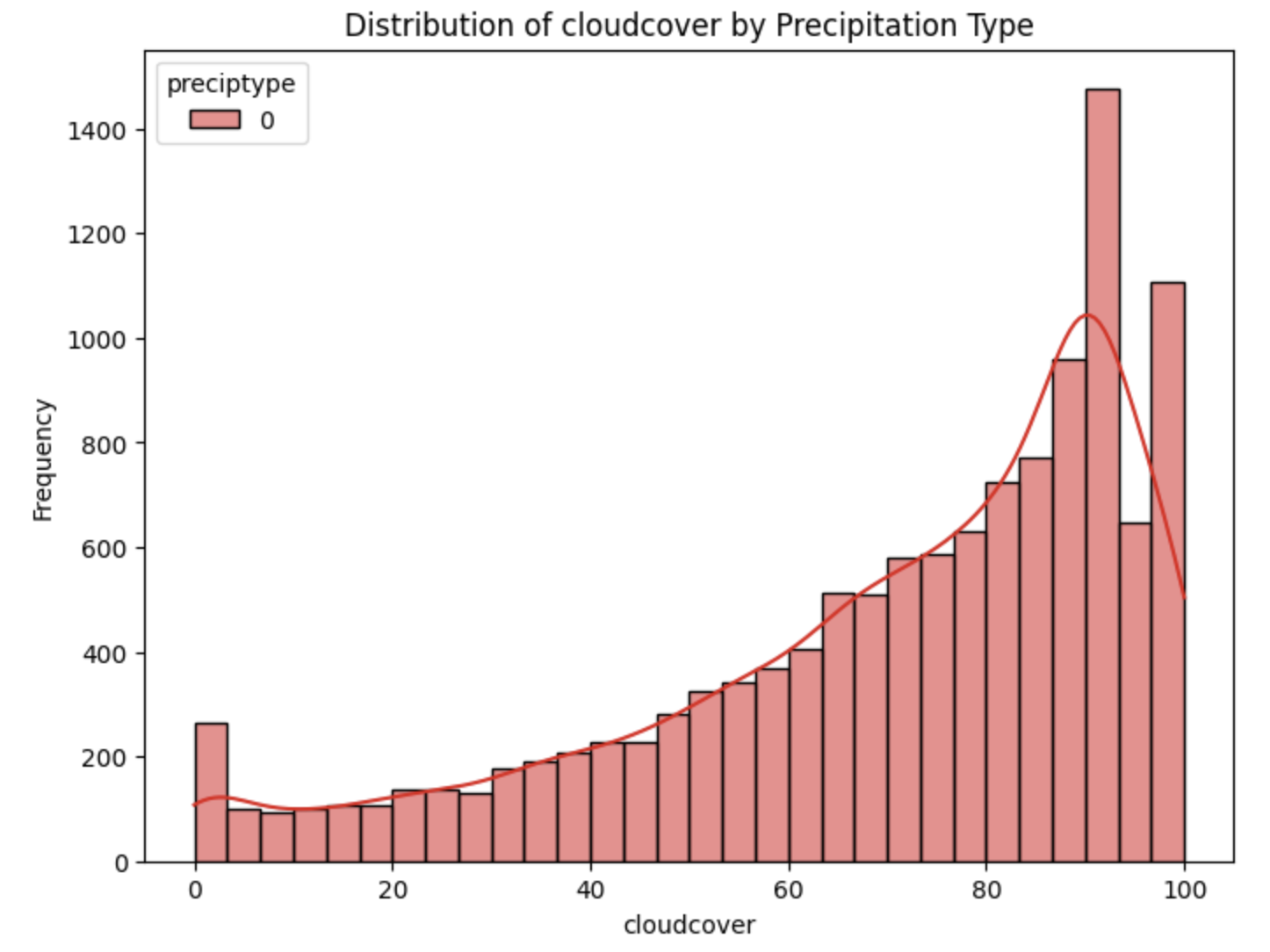


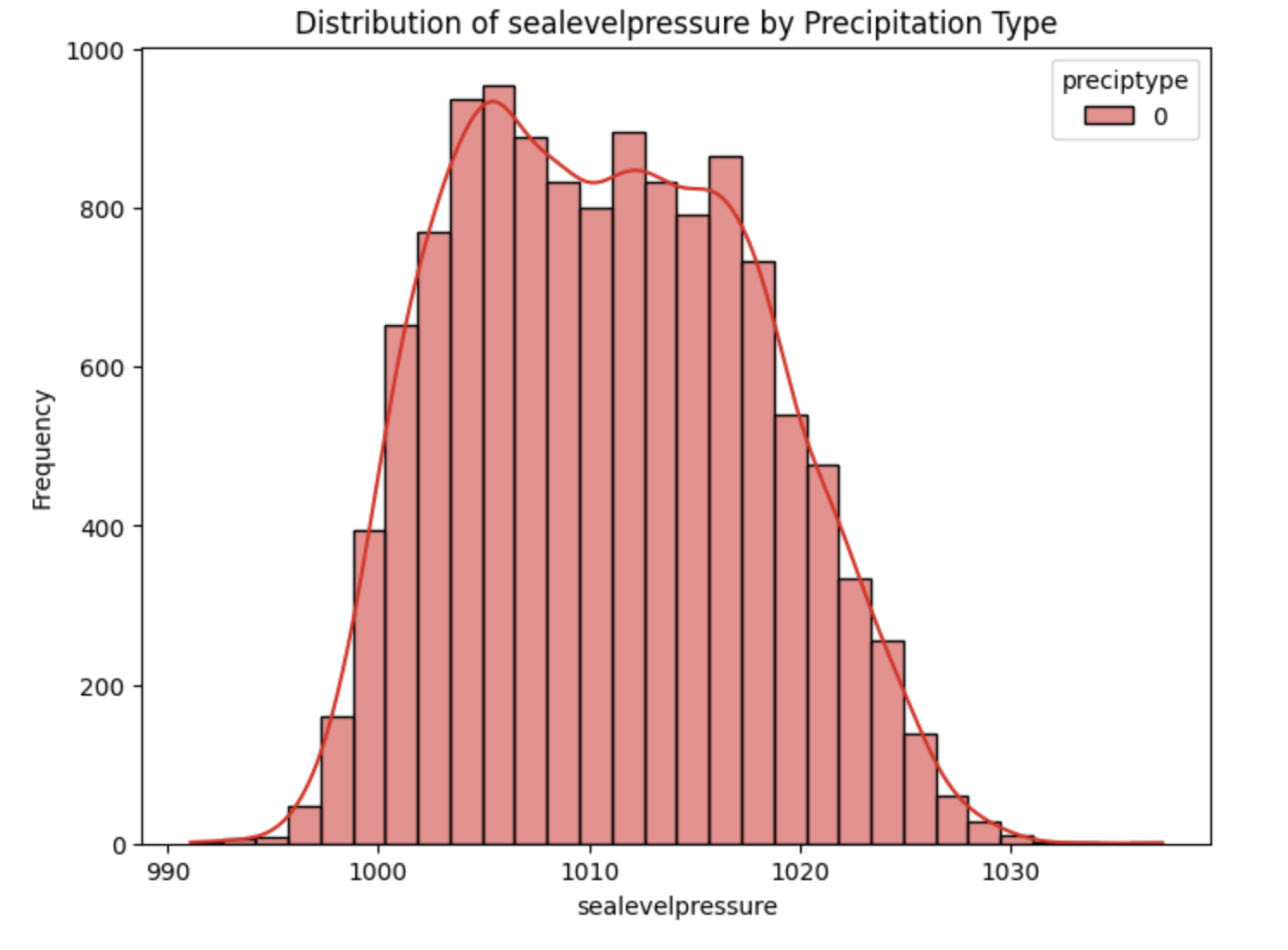
## Biểu đồ phân tích đặc trưng











## Kết luận

### Tổng Quan Dữ Liệu

* **Kích Thước Dữ Liệu**: Tập dữ liệu hiện tại có 12.419 hàng và 13 cột. Số lượng hàng cho thấy đây là một tập dữ liệu lớn, có thể đủ để thực hiện các phân tích thống kê có ý nghĩa.
* **Loại Dữ Liệu**: Các cột trong tập dữ liệu bao gồm các biến số (như nhiệt độ, độ ẩm) và biến phân loại (như điều kiện thời tiết). Điều này cho phép chúng ta thực hiện các phân tích khác nhau

### Chất Lượng Dữ Liệu

* **Thiếu Dữ Liệu**: Tất cả các ô NaN đã được loại bỏ hoặc thay thế bằng giá trị 0 (đối với cột precipt). Điều này đảm bảo rằng không có giá trị thiếu sẽ ảnh hưởng đến kết quả phân tích.
* **Giá Trị Bất Thường**: Đã thực hiện kiểm tra và loại bỏ các giá trị ngoại lệ trong các cột nhiệt độ, giúp dữ liệu trở nên đáng tin cậy hơn cho phân tích.
* **Biến Đã Được Chuyển Đổi**: Cột 'preciptype' đã được dán nhãn với hai giá trị là '0' (không có mưa) và '1' (có mưa). Cột ‘conditions’ được dán nhãn không thứ tự, thể hiện theo sự đặc trưng của từng giá trị.
* **Biểu Đồ Phân Phối**: Nhìn vào biểu đồ, ta thấy nhiệt độ có xu hướng tập trung vào một khoảng nhất định, cho thấy tính đồng nhất trong dữ liệu.
* **Khả Năng Phân Tích**: Với dữ liệu đã được tiền xử lý, ta có thể tiến hành phân tích thống kê để tìm hiểu mối quan hệ giữa các yếu tố như nhiệt độ và độ ẩm đối với khả năng xuất hiện mưa.

# 

# 

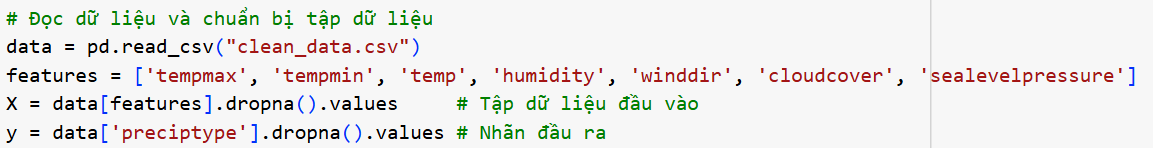
# 

# 

# **4. Xây dựng mô hình và huấn luyện**

## 4.1. Đọc dữ liệu và chọn các trường dùng để huấn luyện mô hình

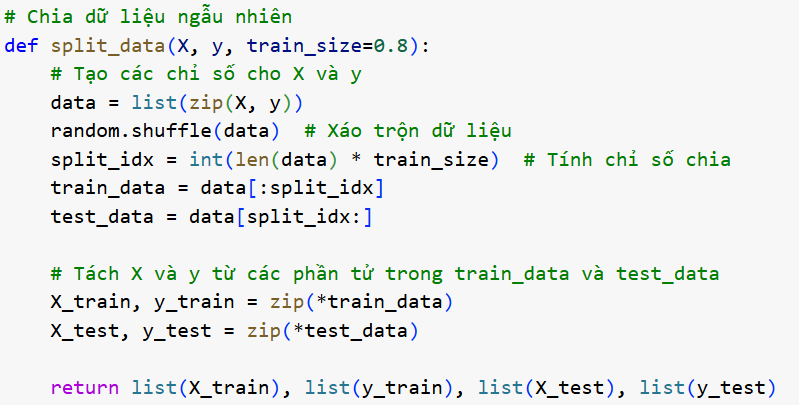
* Nhóm lựa chọn các trường đặc trưng dưới đây để huấn luyện mô hình:
  + tempmax: nhiệt độ cao nhất trong ngày
  + tempmin: nhiệt độ thấp nhất trong ngày
  + temp: nhiệt độ trung bình
  + humidity: độ ẩm
  + winddir: hướng gió
  + cloudcover: độ che phủ mây
  + sealevelpressure: áp suất khí quyển ở mực nước biển
  + preciptype: dùng để xác định có mưa hay không (1: mưa, 0: không mưa)



## 4.2 Chia dữ liệu

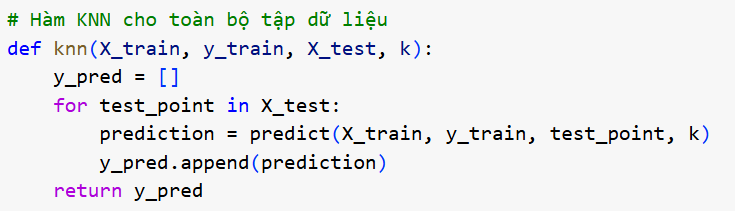
* Nhóm lựa chọn chia dữ liệu thành tập huấn luyện và tập kiểm tra với tỷ lệ 80 - 20.
  + X\_train: Là tập dữ liệu đầu vào (features) dùng để huấn luyện mô hình. Mỗi phần tử trong X\_train là một vector đặc trưng của một mẫu dữ liệu (ví dụ: nhiệt độ, độ ẩm, áp suất, v.v.).
  + y\_train: Là tập nhãn (labels) tương ứng với X\_train. Mỗi phần tử trong y\_train là một giá trị đầu ra (dự báo mưa hoặc không mưa) liên kết với một mẫu dữ liệu trong X\_train.
  + X\_test: Là tập dữ liệu đầu vào (features) dùng để kiểm tra mô hình. Đây là các mẫu dữ liệu mà mô hình chưa được "thấy" trước đó trong quá trình huấn luyện.
  + y\_test: Là tập nhãn (labels) tương ứng với X\_test. y\_test chứa các giá trị thực tế của dữ liệu kiểm tra, dùng để so sánh với kết quả dự đoán (y\_pred) từ mô hình, nhằm đánh giá độ chính xác (accuracy) của mô hình.



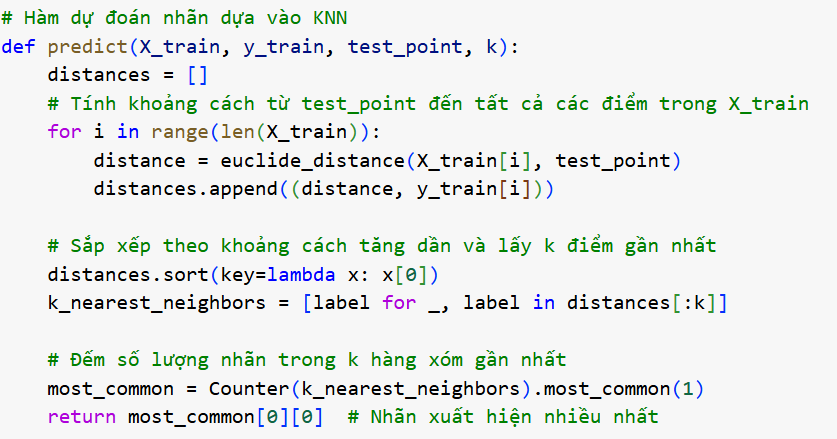


## 4.3 Lập trình mô hình

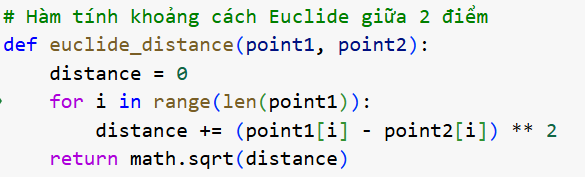
* Áp dụng thuật toán KNN cho toàn bộ tập dữ liệu kiểm tra (X\_test). Nó sử dụng hàm ‘predict’ để dự đoán nhãn cho từng điểm dữ liệu thử nghiệm trong tập kiểm tra.



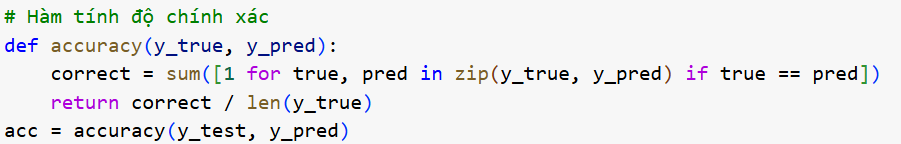
* Hàm ‘predict’ thực hiện dự đoán nhãn (label) cho một điểm dữ liệu thử nghiệm (test\_point) dựa trên thuật toán KNN. Nó sử dụng tập dữ liệu huấn luyện để tìm k điểm gần nhất với test\_point và đưa ra nhãn dự đoán dựa trên nhãn xuất hiện nhiều nhất trong k hàng xóm gần nhất.



* Tính khoảng cách Euclide giữa hai điểm trong không gian nhiều chiều. Đây là một phương pháp đo khoảng cách phổ biến trong các thuật toán KNN để xác định độ tương đồng giữa các điểm dữ liệu.

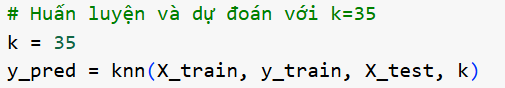


* Hàm accuracy được sử dụng để đánh giá hiệu suất của mô hình dự đoán. Cụ thể, nó tính toán độ chính xác (accuracy) của mô hình bằng cách đo tỷ lệ dự đoán đúng (giá trị dự đoán khớp với giá trị thực tế) trên tổng số mẫu dữ liệu.

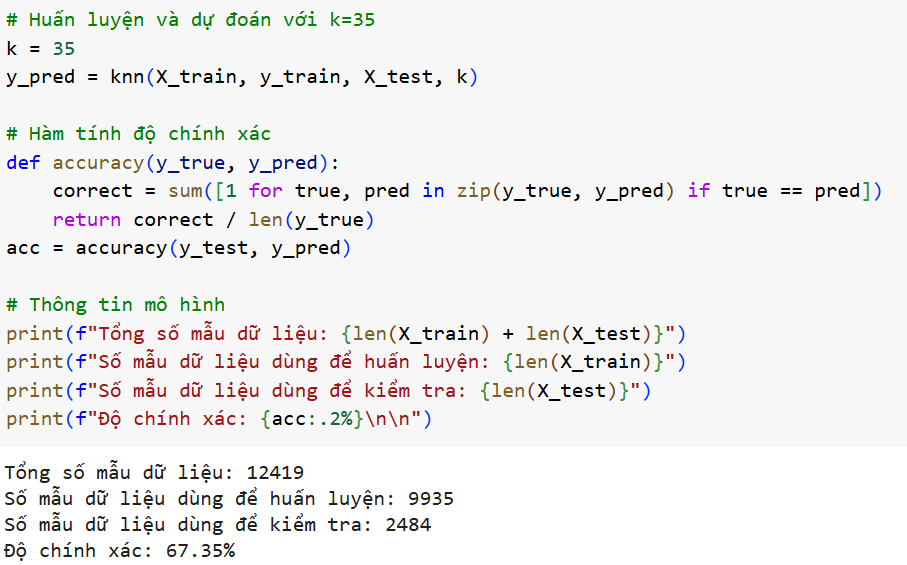


## 4.4 Huấn luyện và kiểm tra mô hình

* Thử với k = 35:



* Kết quả:



# **5. Đánh giá mô hình**

## 5.1. Đánh giá về thuật toán được lựa chọn K-NN

### a. Ưu điểm của thuật toán K-NN trong dự án

* **Dễ triển khai và hiểu rõ**: K-NN là thuật toán đơn giản, không yêu cầu nhiều giả định về dữ liệu.
* **Không cần huấn luyện:** K-NN là thuật toán lazy learning, không cần giai đoạn huấn luyện phức tạp, chỉ cần tính khoảng cách khi dự đoán.
* **Hiệu quả cho dữ liệu nhỏ và vừa:** Với tập dữ liệu vừa phải như dự án này, K-NN hoạt động tốt mà không tốn nhiều tài nguyên tính toán.
* **Linh hoạt với các loại dữ liệu**: K-NN có thể áp dụng cho dữ liệu phân loại (như dự đoán "mưa" hay "không mưa") và dữ liệu hồi quy.

### b. Nhược điểm của thuật toán K-NN

* **Nhạy cảm với lựa chọn k:** Giá trị k ảnh hưởng mạnh đến độ chính xác. Chọn k nhỏ sẽ làm tăng độ nhiễu, chọn k lớn có thể làm mất đi các đặc trưng cục bộ.
* **Tính toán chậm với dữ liệu lớn:** Thuật toán phải tính khoảng cách với mọi điểm trong tập huấn luyện, gây tốn thời gian khi dữ liệu lớn.
* **Nhạy cảm với outliers (điểm ngoại lai):** Các điểm bất thường có thể ảnh hưởng lớn đến kết quả dự đoán.
* **Không thể tự động tối ưu hóa:** Không có cơ chế tự học; giá trị k phải được lựa chọn thủ công qua thử nghiệm.

## 5.2. Đánh giá về sự thay đổi của k và độ chính xác (accuracy)

* Tạo các dữ liệu để đánh giá ngay khi thực hiện việc huấn luyện mô hình:

Sử dụng dữ liệu đã lấy theo tỷ lệ training-testing là 80% - 20% để đánh giá và kiểm thử mô hình.

Dữ liệu huấn luyện / dữ liệu kiểm tra: 9935 / 2484.

* Kết quả chạy thử nghiệm mô hình với các khoảng k:



**Quan sát từ biểu đồ:**

* Khi giá trị k nhỏ (k=25), độ chính xác của mô hình đạt khoảng 0.678.
* Khi tăng dần k lên đến 55, độ chính xác tăng nhẹ và **đạt đỉnh tại k=55 (~0.682)**.
* Khi k tiếp tục tăng lên 65, 75, 85...115, độ chính xác giảm dần.

**Giải thích lý do:**

* K nhỏ (Underfitting - chưa khái quát hóa):

Khi k quá nhỏ, thuật toán dựa vào rất ít điểm lân cận để dự đoán. Điều này làm cho mô hình nhạy cảm với nhiễu (outliers), dẫn đến độ chính xác không ổn định.

Mô hình có thể bị overfitting với các mẫu cục bộ.

* K lớn (Overfitting - mất chi tiết cục bộ):

Khi k tăng, K-NN xem xét nhiều hàng xóm hơn. Điều này giúp giảm nhiễu, nhưng có thể làm mất đi các đặc trưng cục bộ.

Nếu k quá lớn, mọi điểm sẽ "hòa trộn" và dự đoán bị chi phối bởi nhóm lớn hơn, làm giảm độ chính xác.

* Giá trị k tối ưu:

Độ chính xác đạt cao nhất ở **khoảng** k=55.

Lý do: Tại **khoảng** k này, thuật toán cân bằng giữa việc giảm nhiễu và bảo toàn thông tin cục bộ.

## 5.3.Các biểu đồ đánh giá mô hình (với k =55)

## 

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

* Ma trận nhầm lẫn trong hình được tổ chức như sau:

|  | Predicted: No Rain | Predicted: Rain |
| --- | --- | --- |
| Actual: No Rain | 878 (TN) | 466 (FP) |
| Actual: Rain | 330 (FN) | 810 (TP) |

TN (True Negative): 878

Số lượng trường hợp thực tế không có mưa và mô hình cũng dự đoán không có mưa.

FP (False Positive): 466

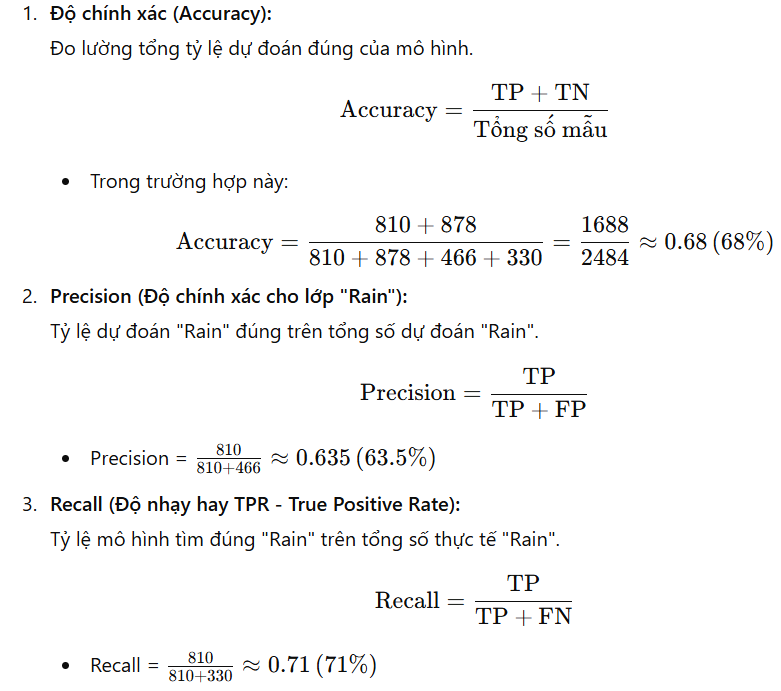
Số lượng trường hợp thực tế không có mưa nhưng mô hình dự đoán có mưa.

FN (False Negative): 330

Số lượng trường hợp thực tế có mưa nhưng mô hình dự đoán không có mưa.

TP (True Positive): 810

Số lượng trường hợp thực tế có mưa và mô hình dự đoán đúng là có mưa.

* Từ ma trận nhầm lẫn, ta có thể tính toán các chỉ số quan trọng sau:
* 

# **6. Kết luận**

Mô hình dự báo mưa sử dụng thuật toán K-Nearest Neighbors (KNN) đã cho kết quả khả quan với độ chính xác khoảng 68% khi giá trị k tối ưu là 55. KNN là một thuật toán đơn giản, dễ triển khai và không yêu cầu quá trình huấn luyện phức tạp, rất phù hợp với các tập dữ liệu nhỏ và vừa như trong nghiên cứu này. Tuy nhiên, mô hình cũng có một số hạn chế, đặc biệt là khi đối mặt với các giá trị k không tối ưu, gây ra hiện tượng underfitting hoặc overfitting, làm giảm độ chính xác của dự báo.

Mặc dù KNN có thể không phải là mô hình tối ưu nhất trong mọi tình huống, nhưng với các cải tiến như tối ưu hóa giá trị k, loại bỏ các điểm ngoại lai và kết hợp với các phương pháp tiền xử lý dữ liệu, mô hình này vẫn có thể đạt được hiệu suất tốt hơn trong các bài toán dự báo mưa. Việc đánh giá mô hình liên tục và thử nghiệm trên các bộ dữ liệu khác sẽ giúp cải thiện khả năng tổng quát và độ chính xác của mô hình.

* Nhận xét:  
  Biểu đồ của các giá trị mục tiêu quan sát được so với các giá trị dự báo để kiểm tra xem các động lực nhất định của dữ liệu mà mô hình thu thập được.

Sai số phần trăm tuyệt đối trung bình (MAPE) giữa giá trị thực tế và giá trị dự báo.

Lỗi bình phương trung bình gốc (RMSE), có thể với sự chuẩn hóa, giữa các giá trị thực tế và dự báo.

Sai số tuyệt đối trung bình (MAE), có thể với sự chuẩn hóa, giữa các giá trị thực tế và dự báo.