|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI



BÁO CÁO BÀI TẬP LỚN HỌC PHẦN

DỮ LIỆU LỚN

Đề tài: dự đoán xác suất tỷ lệ có MƯA HAY không

**Giáo viên hướng dẫn: TS TẠ QUANG CHIỂU**

Nhóm sinh viên thực hiện: NHÓM 4

1. Bùi Tuấn Kiên, Lớp 64HTTT3
2. Trần Thu Hương, Lớp 64HTTT3
3. Đào Công Minh Vũ, Lớp 64HTTT3

Hà Nội, tháng 10 năm 2024

LỜI Cảm ƠN

Để hoàn thành được bài tập lớn này, chúng em xin gửi lời cảm ơn đến thầy Tạ Quang Chiểu đã tận tình hướng dẫn trong suốt quá trình học tập.

Mặc dù đã có nhiều cố gắng trong suốt quá trình làm bài tập, nhưng kiến thức còn hạn chế, cũng như thời gian có hạn nên không thể tránh khỏi những sai sót. Vì vậy, nhóm em mong thầy góp ý để có thể chỉnh sửa, hoàn thiện bài làm tốt hơn.

Trân trọng cảm ơn thầy!

.

**Lời mở đầu.................................................................................................................4**

**Phần 1. Tổng quan.........................................................................................................5**

1. **Giới thiệu về học máy................................................................................ 5**
2. **Trình bày các phương pháp học máy được sử dụng trong đề tài nhóm**

**chọn................................................................................................................. 8 Thuật Toán Logistic….............................................................................8**

* + - **Input: .......................................................................................................8**
    - **Output:....................................................................................................8**
    - **Hàm mất mát của phương pháp:..........................................................8**
    - **Cách giải bài toán tối ưu hoặc các bước thực hiện của thuật toán....9**
    - **Các độ đo để đánh giá chất lượng mô hình dự đoán.......................... 9**

**Phần 2. Thực nghiệm.....................................................................................................9**

1. **Mô tả bài toán...........................................................................................9**
2. **Mô tả tập dữ liệu của bài toán................................................................ 11**
3. **Viết ứng dụng........................................................................................... 11**
4. **Phân tích kết quả của chương trình.......................................................17**

**Phần 3. Kết luận.................................................................................................19**

**Tài liệu tham khảo.............................................................................................19**

## Lời mở đầu

**Học máy** [(tiếng](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh) [Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh) : *machine learning*) là một lĩnh vực của [trí](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o) [tuệ](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o) [nhân](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o) [tạo](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o) liên quan đến việc nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể.

Học máy có liên quan lớn đến [thống](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA) [kê](https://vi.wikipedia.org/wiki/Khoa_h%E1%BB%8Dc_Th%E1%BB%91ng_k%C3%AA), vì cả hai lĩnh vực đều nghiên cứu việc phân tích dữ liệu, nhưng khác với thống kê, học máy tập trung vào sự phức tạp của các giải thuật trong việc thực thi tính toán. Nhiều bài toán suy luận được xếp vào loại bài toán [NP-khó](https://vi.wikipedia.org/wiki/NP-kh%C3%B3), vì thế một phần của học máy là nghiên cứu sự phát triển các giải thuật suy luận xấp xỉ mà có thể xử lý được.

Học máy có hiện nay được áp dụng rộng rãi bao gồm [máy](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u) [truy](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u) [tìm](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u) [dữ](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u) [liệu](https://vi.wikipedia.org/wiki/M%C3%A1y_truy_t%C3%ACm_d%E1%BB%AF_li%E1%BB%87u)[, chẩn](https://vi.wikipedia.org/wiki/Ch%E1%BA%A9n_%C4%91o%C3%A1n) [đoán](https://vi.wikipedia.org/wiki/Ch%E1%BA%A9n_%C4%91o%C3%A1n) [y](https://vi.wikipedia.org/wiki/Ch%E1%BA%A9n_%C4%91o%C3%A1n) [khoa](https://vi.wikipedia.org/wiki/Ch%E1%BA%A9n_%C4%91o%C3%A1n), phát hiện [thẻ](https://vi.wikipedia.org/w/index.php?title=Th%E1%BA%BB_t%C3%ADn_d%E1%BB%A5ng_gi%E1%BA%A3&action=edit&redlink=1) [tín](https://vi.wikipedia.org/w/index.php?title=Th%E1%BA%BB_t%C3%ADn_d%E1%BB%A5ng_gi%E1%BA%A3&action=edit&redlink=1) [dụng](https://vi.wikipedia.org/w/index.php?title=Th%E1%BA%BB_t%C3%ADn_d%E1%BB%A5ng_gi%E1%BA%A3&action=edit&redlink=1) [giả](https://vi.wikipedia.org/w/index.php?title=Th%E1%BA%BB_t%C3%ADn_d%E1%BB%A5ng_gi%E1%BA%A3&action=edit&redlink=1), phân tích [thị](https://vi.wikipedia.org/wiki/Th%E1%BB%8B_tr%C6%B0%E1%BB%9Dng_ch%E1%BB%A9ng_kho%C3%A1n) [trường](https://vi.wikipedia.org/wiki/Th%E1%BB%8B_tr%C6%B0%E1%BB%9Dng_ch%E1%BB%A9ng_kho%C3%A1n) [chứng](https://vi.wikipedia.org/wiki/Th%E1%BB%8B_tr%C6%B0%E1%BB%9Dng_ch%E1%BB%A9ng_kho%C3%A1n) [khoán](https://vi.wikipedia.org/wiki/Th%E1%BB%8B_tr%C6%B0%E1%BB%9Dng_ch%E1%BB%A9ng_kho%C3%A1n), phân loại các [chuỗi](https://vi.wikipedia.org/w/index.php?title=Chu%E1%BB%97i_DNA&action=edit&redlink=1) [DNA](https://vi.wikipedia.org/w/index.php?title=Chu%E1%BB%97i_DNA&action=edit&redlink=1)[,](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ti%E1%BA%BFng_n%C3%B3i) [nhận](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ti%E1%BA%BFng_n%C3%B3i) [dạng](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ti%E1%BA%BFng_n%C3%B3i) [tiếng](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ti%E1%BA%BFng_n%C3%B3i) [nói](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ti%E1%BA%BFng_n%C3%B3i) và [chữ](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ch%E1%BB%AF_vi%E1%BA%BFt_tay) [viết](https://vi.wikipedia.org/wiki/Nh%E1%BA%ADn_d%E1%BA%A1ng_ch%E1%BB%AF_vi%E1%BA%BFt_tay)[,](https://vi.wikipedia.org/wiki/D%E1%BB%8Bch_t%E1%BB%B1_%C4%91%E1%BB%99ng) [dịch](https://vi.wikipedia.org/wiki/D%E1%BB%8Bch_t%E1%BB%B1_%C4%91%E1%BB%99ng) [tự](https://vi.wikipedia.org/wiki/D%E1%BB%8Bch_t%E1%BB%B1_%C4%91%E1%BB%99ng) [động](https://vi.wikipedia.org/wiki/D%E1%BB%8Bch_t%E1%BB%B1_%C4%91%E1%BB%99ng)…

Những năm gần đây, sự phát triển của các hệ thống tính toán cùng với lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn đã giúp machine learning tiến thêm một bước dài.

Vì vậy, chúng em đã áp dụng ba thuật toán trong học máy - thuật toán để làm bài phương pháp hồi quy tuyến tính (Logistic Regression) để dự đoán xác suất tỉ lệ mưa hay không.

### Phần 1. Tổng quan

### 1. Giới thiệu về học máy

Machine Learning, hay Học Máy, là một lĩnh vực của trí tuệ nhân tạo (Artificial Intelligence – AI), cho phép máy tính tự động học hỏi và cải thiện hiệu suất của nó trong việc thực hiện các nhiệm vụ cụ thể mà không cần sự can thiệp của con người.

**● Lịch sử của Machine Learning.**

* **Thập kỷ 1940-1950:** Thời kỳ đầu của máy tính và ý tưởng đầu tiên về máy học:

+ Trong thời kỳ này, máy tính đã xuất hiện và người ta đã bắt đầu nghĩ đến khả năng sử dụng chúng để mô phỏng khả năng học của con người.

+ Alan Turing đã đưa ra khái niệm về máy Turing, một ý tưởng quan trọng trong phát triển machine learning.

* **Thập kỷ 1950-1960:** Mạng nơ-ron và học máy cổ điển:

+ Warren McCulloch và Walter Pitts đã giới thiệu mô hình nơron đầu tiên, tạo cơ sở cho phát triển mạng nơ-ron.

+ Frank Rosenblatt phát triển Perceptron, một mô hình học máy đơn giản được xem là tiền đề của các mô hình neural network ngày nay.

* **Thập kỷ 1960-1970:** Sự gián đoạn và suy giảm động lực:

+ Nhiều nghiên cứu cho thấy giới hạn của Perceptron, khiến nhiều người nghi ngờ về khả năng của machine learning.

* **Thập kỷ 1980-1990:** Sự tái sinh và sự xuất hiện của học máy thống kê:

+ Xuất hiện các phương pháp thống kê trong học máy, chẳng hạn như Support Vector Machines (SVM) và Decision Trees.

+ Sự phát triển của khái niệm "học máy thống kê" giúp tăng cường độ tin cậy của machine learning.

* **Thập kỷ 1990-2000:** Mạng nơ-ron tái xuất và sự phổ biến của học sâu:

+ Sự gia tăng của dữ liệu và sức mạnh tính toán đã tạo điều kiện cho việc phát triển mạnh mẽ của mạng nơ-ron đa tầng (Multilayer Perceptron) và học sâu.

+ Xuất hiện nhiều thuật toán machine learning mới và sự chú ý tăng lên đối với lĩnh vực này.

* **Thập kỷ 2010-nay:** Cuộc cách mạng của học máy và trí tuệ nhân tạo:

+ Tiến bộ lớn trong deep learning, đặc biệt là Convolutional Neural Networks (CNN) và Recurrent Neural Networks (RNN).

+ Sự phổ biến của các ứng dụng machine learning trong nhiều lĩnh vực như xử lý ngôn ngữ tự nhiên, thị giác máy tính, và tự lái xe ô tô.

+ Sự gia tăng của phương pháp học máy không giám sát và học máy tăng cường.

* **Vai trò của Machine Learning: *Machine Learning*** đóng vai trò quan trọng trong nhiều lĩnh vực và mang lại nhiều lợi ích. Dưới đây là 1 số vai trò chính:
  + Dự đoán và dự báo: Machine learning có khả năng học từ dữ liệu quá khứ để dự đoán và dự báo các sự kiện tương lai.
  + Phân loại và nhận dạng: Machine learning có khả năng phân loại và nhận dạng các đối tượng và mẫu dữ liệu.
  + Tư vấn và đề xuất: Machine learning có khả năng tư vấn và đề xuất dựa trên thông tin và hành vi người dùng.
  + Xử lý và phân tích dữ liệu lớn: Machine learning giúp xử lý và phân tích dữ liệu lớn (big data) để tìm ra các mẫu, xu hướng và thông tin hữu ích.
  + Tự động hóa và tối ưu hóa: Machine learning có thể tự động hóa quy trình và tối ưu hóa các hệ thống và quy trình phức tạp.
* Hiểu và xử lý ngôn ngữ tự nhiên: Machine learning giúp máy tính hiểu và xử lý ngôn ngữ tự nhiên, từ việc nhận dạng và tự động sinh văn bản đến dịch máy và phân tích ý kiến.
* **Trình bày Học có giám sát, Học không có giám sát**

|  |  |  |  |
| --- | --- | --- | --- |
| Kiểu học | Tập dữ liệu | Bài toán | Phương pháp |
| Học có giám  sát | Tập dữ liệu huấn luyện được  gán nhãn (Xtrain, ytrain) | Hồi quy, phân lớp | Linear Regression,  Lasso, Ridge,  SVM, Perceptron,  ID3, Neural  Network, Logistic  Regression |
| Học không có giám sát | Tập dữ liệu huấn luyện không được gán nhãn (Xtrain) | Phân cụm | K-means |

* **Bài toán Hồi quy, bài toán Phân lớp, bài toán Phân cụm.**

|  |  |  |
| --- | --- | --- |
| Bài toán | Tập dữ liệu | Các phương pháp |
| Hồi quy | Tập dữ liệu huấn luyện đã được gán nhãn (Xtrain, ytrain). Miền giá trị của tập nhãn ytrain là liên tục và có vô số giá trị | Linear Regression, Lasso,  Ridge, |
| Phân lớp | Tập dữ liệu huấn luyện đã được gán nhãn (Xtrain, ytrain). Miền giá trị của tập nhãn ytrain là rời rạc và hữu hạn | SVM, perceptron, cây quyết định, Neural Network, Logistic  Regression |
| Phân Cụm | Tập dữ liệu huấn luyện không  được gán nhãn (Xtrain,) | K-means |

**2. Trình bày phương pháp học máy được sử dụng trong đề tài của nhóm**

**Thuật Toán Logistic Regression:**

**- Khái niệm:** Logistic Regression là một phương pháp thống kê được sử dụng để phân loại và hồi quy, đặc biệt hữu ích trong các bài toán phân loại nhị phân (binary classification). Thuật toán này ước lượng xác suất mà một đối tượng thuộc về một trong hai lớp thông qua một hàm logistic (hay sigmoid)

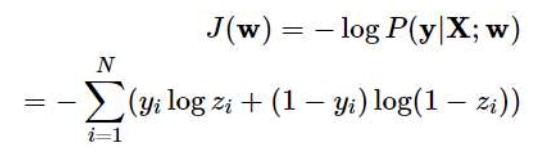
* **Input:** Tập dữ liệu huấn luyện đã được gán nhãn (Xtrain,ytrain)
* **Output:** Được thể hiện dưới dạng xác suất (probability)

Đầu ra dự đoán: 

* **Hàm mất mát của phương pháp:**



Hoặc:

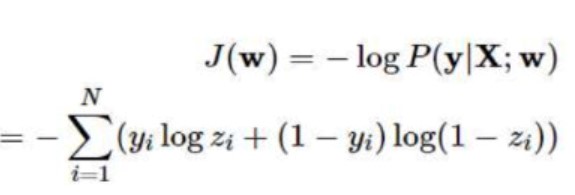


* **Cách giải bài toán tối ưu hoặc các bước thực hiện của thuật toán**

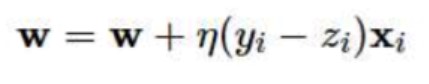
+ Trực tiếp tối ưu hàm số này theo w khó

+ Một phương pháp thường được sử dụng đó là lấy logarit tự nhiên (cơ số e) của likelihood function biến phép nhân thành phép cộng và để tránh việc số quá nhỏ

+ Sau đó lấy ngược dấu để một hàm và coi nó là hàm mất mát. Lúc này bài toán tìm giá trị lớn nhất (maximum likelihood) trở thành bài toán tìm giá trị nhỏ nhất của hàm mất mát (hàm này còn được gọi là negative log likelihood):



+ Chúng ta sử dụng **phương pháp Gradient Descent** để tìm w



* **Các độ đo để đánh giá chất lượng mô hình dự đoán.**

+ Accuracy

+ Precision

+ Recall

+ F1\_score

### Phần 2. Thực nghiệm

#### **1. Mô tả bài toán**

* **Tên bài toán:** Dự đoán xác suất tỉ lệ có mưa hay không
* **Mục đích của bài toán:**

+ Mục tiêu của bài toán "Dự đoán xác suất tỉ lệ có mưa hay không" là xây dựng một mô hình học máy có khả năng xác định thời tiết có mưa hay không mưa. Mô hình này sẽ được huấn luyện trên tập dữ liệu bao gồm các thuộc tính khí tượng học như độ ẩm, nhiệt độ, tốc độ gió, hướng gió, tầm nhìn, áp suất.

+ Việc dự đoán xác suất mưa sẽ hỗ trợ các chuyên gia khí tượng, nhà nghiên cứu và cộng đồng trong việc đưa ra quyết định sớm về thời tiết. Điều này giúp cải thiện việc chuẩn bị cho các tình huống như du lịch, nông nghiệp, và các hoạt động ngoài trời khác, đồng thời tăng tính an toàn và hiệu quả.

* **Input bài toán:** Để dự đoán xác suất tỉ lệ có mưa hay không, các thông tin đầu vào có thể bao gồm:

+ Temperature (Nhiệt độ), Humidity (Độ ẩm), Wind Speed (Tốc độ gió), Wind Bearing (Hướng gió), Visibility (Tầm nhìn), Pressure Millibars (Áp suất)

* **Output bài toán**: Đưa ra dự đoán là có mưa hay không mưa
* **Tóm tắt công việc thực hiện của bài toán:**

+ Bài toán dựa vào dữ liệu có sẵn (đọc dữ liệu từ file weather.csv được lấy trên kaggle ) gồm các trường thông tin **Temperature (Nhiệt độ), Humidity (Độ ẩm), Wind Speed (Tốc độ gió), Wind Bearing (Hướng gió), Visibility (Tầm nhìn), Pressure Millibars (Áp suất)**

+ Tạo ra train set và test set đưa vào huấn luyện tìm ra quy luật. (chia dữ liệu thành hai tập dt\_test và dt\_train trong đó tập dt\_test lấy ra 20%)

+ Từ đó dự đoán ra **có mưa hay không mưa** dựa vào **Temperature (Nhiệt độ), Humidity (Độ ẩm), Wind Speed (Tốc độ gió), Wind Bearing (Hướng gió), Visibility (Tầm nhìn), Pressure Millibars (Áp suất)**

+ Tính hiệu suất Accuracy (so sánh kết quả dự đoán và dữ liệu test) và đưa ra mô hình tốt nhất

#### **2. Mô tả tập dữ liệu của bài toán**

* **Dữ liệu gồm những chiều thông tin:**

+ Temperature (Nhiệt độ), Humidity (Độ ẩm), Wind Speed (Tốc độ gió), Wind Bearing (Hướng gió), Visibility (Tầm nhìn), Pressure Millibars (Áp suất) **-Có 10001 mẫu dữ liệu.**

* **Mô tả nhãn lớp của dữ liệu:**

Ma trận dữ liệu (X) gồm 6 thông tin:

+ Temperature: Nhiệt độ

+ Humidity : Độ ẩm

+ Wind Speed: Tốc độ gió

+ Wind Bearing: Hướng gió

+ Visibility : Tầm nhìn

+ Pressure Millibars: Áp suất

Nhãn lớp (Y) gồm 1 thông tin:

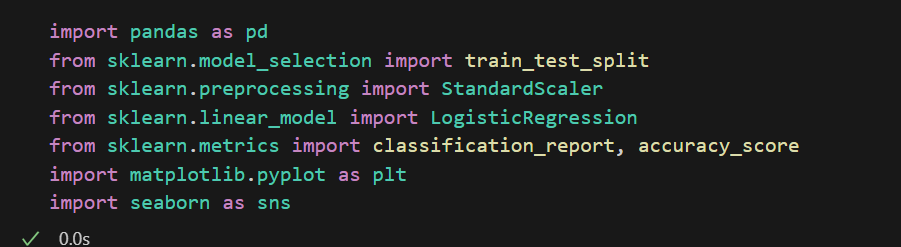
+ Rain: 0 (Không mưa), 1 (Có mưa)

**- Chia tập dữ liệu thành 2 phần: 80% dùng để huấn luyện mô hình, 20% dùng để kiểm tra sự phù hợp của mô hình.**

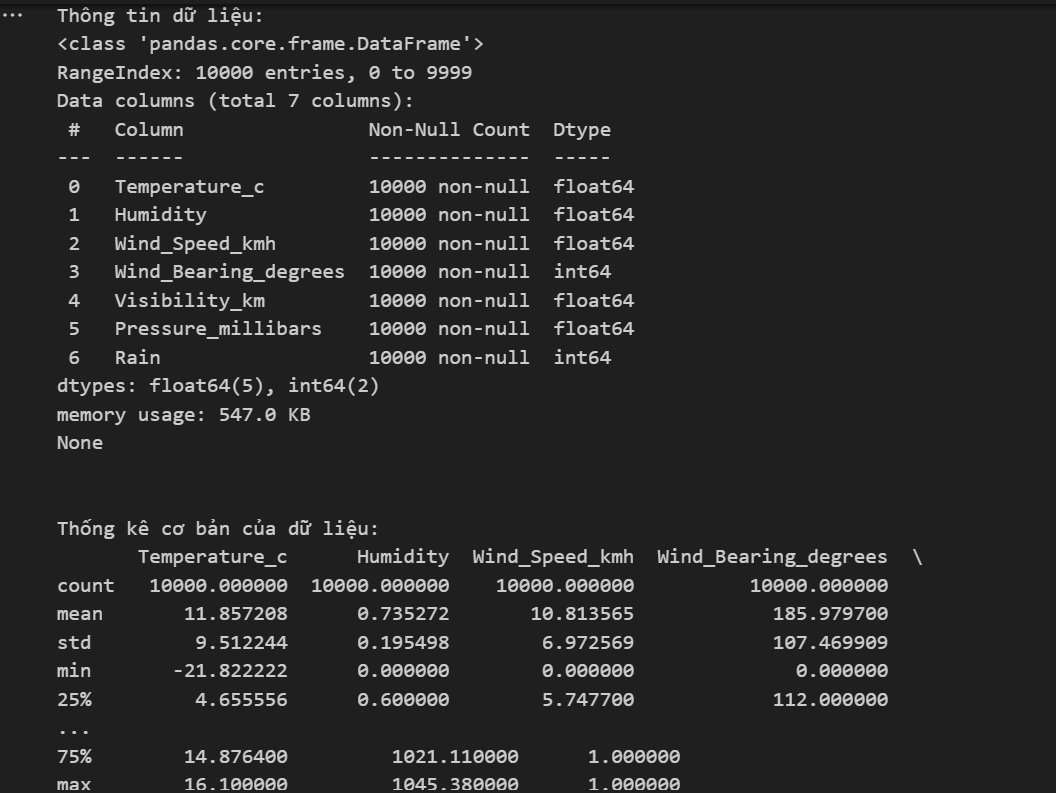
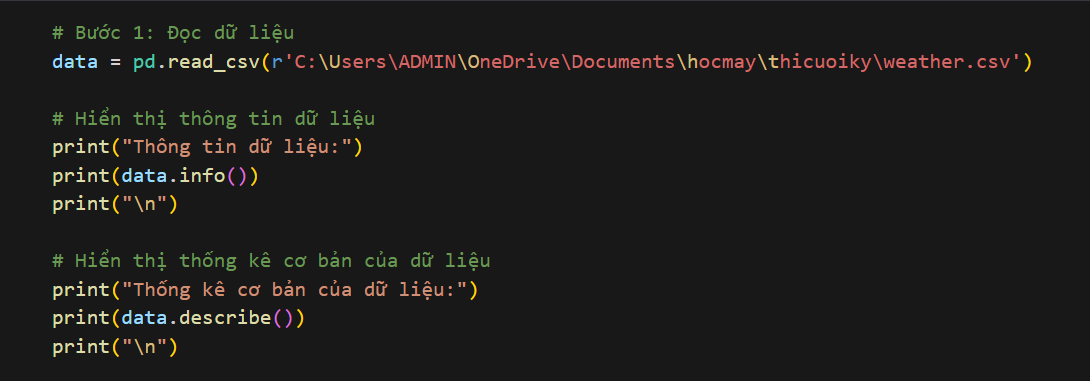
#### **3. Viết ứng dụng**

**a) Cách tiền xử lý dữ liệu**

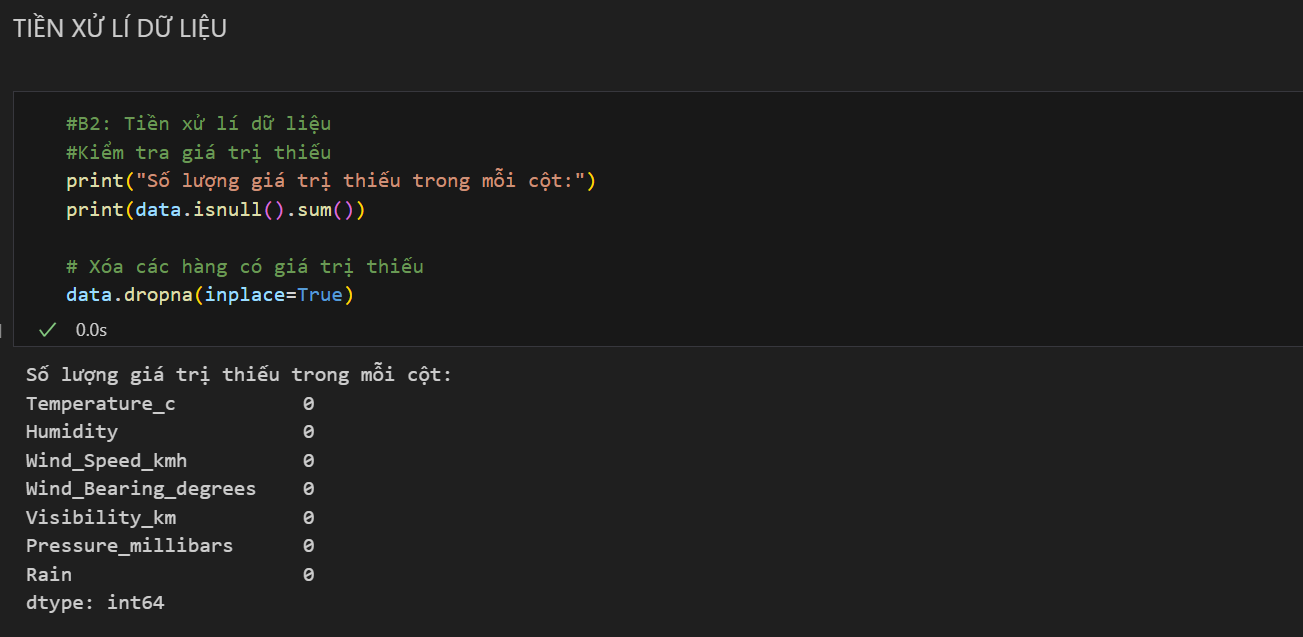
**Import thư viện**



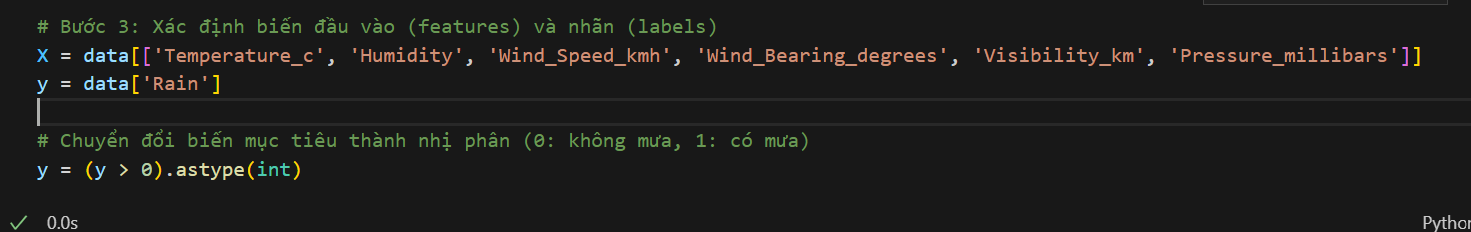
**B1: Đọc dữ liệu và hiển thị thông tin dữ liệu**



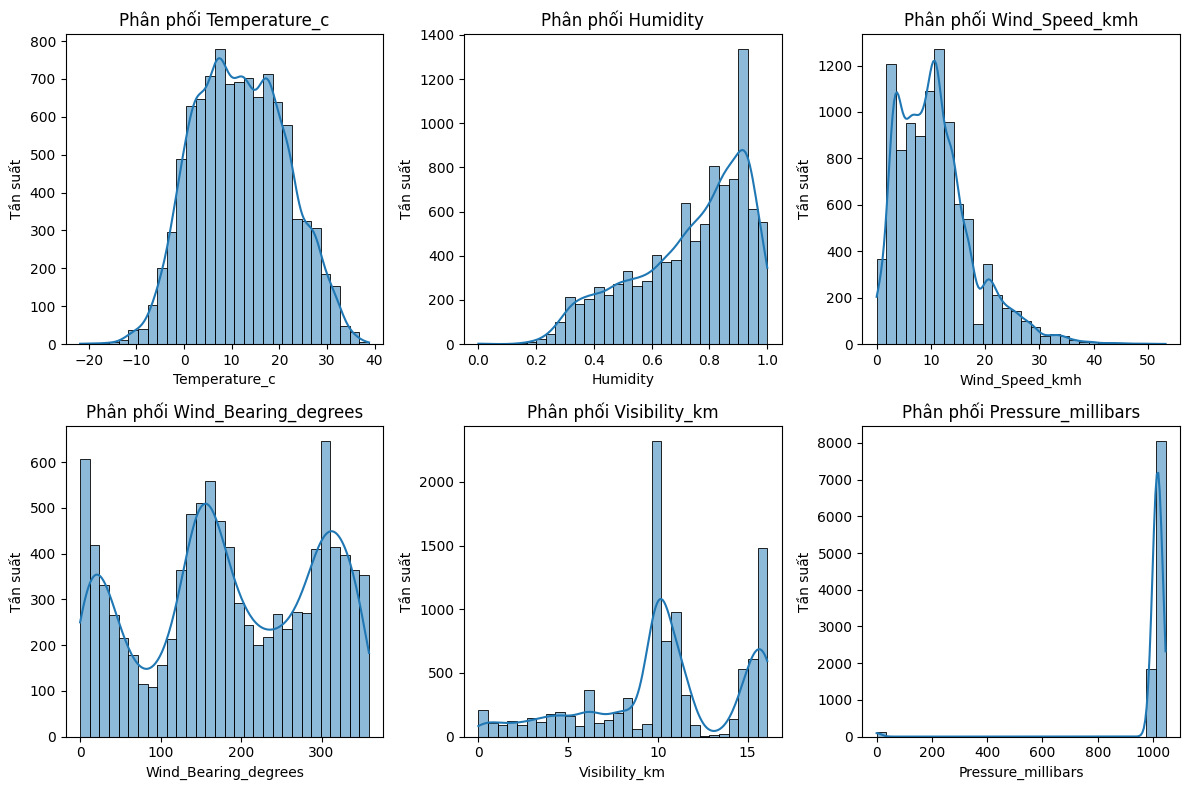
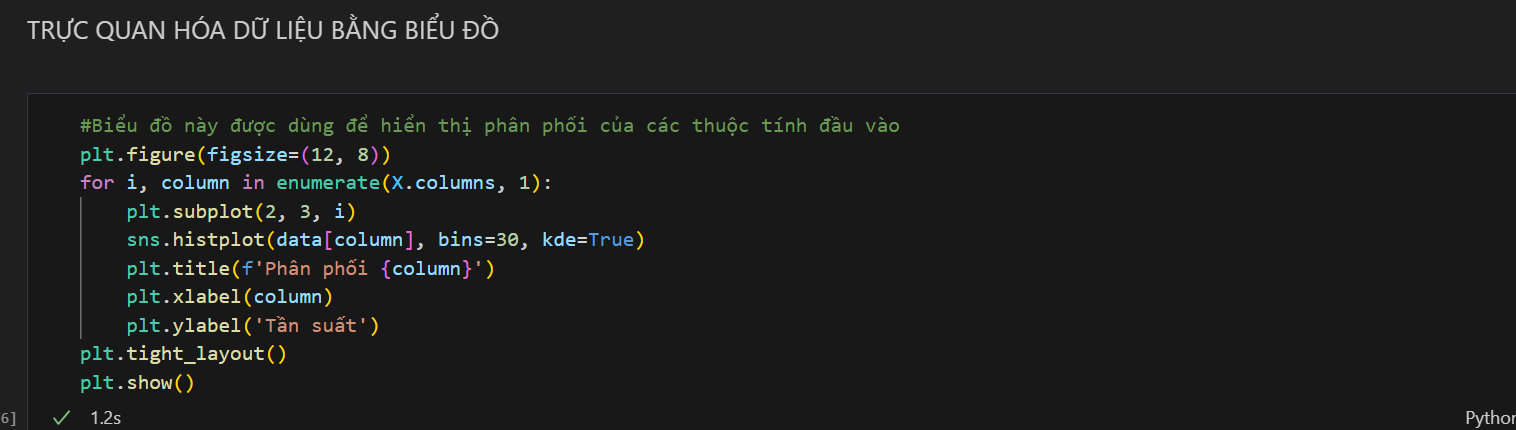
B2: Tiền xử lí dữ liệu



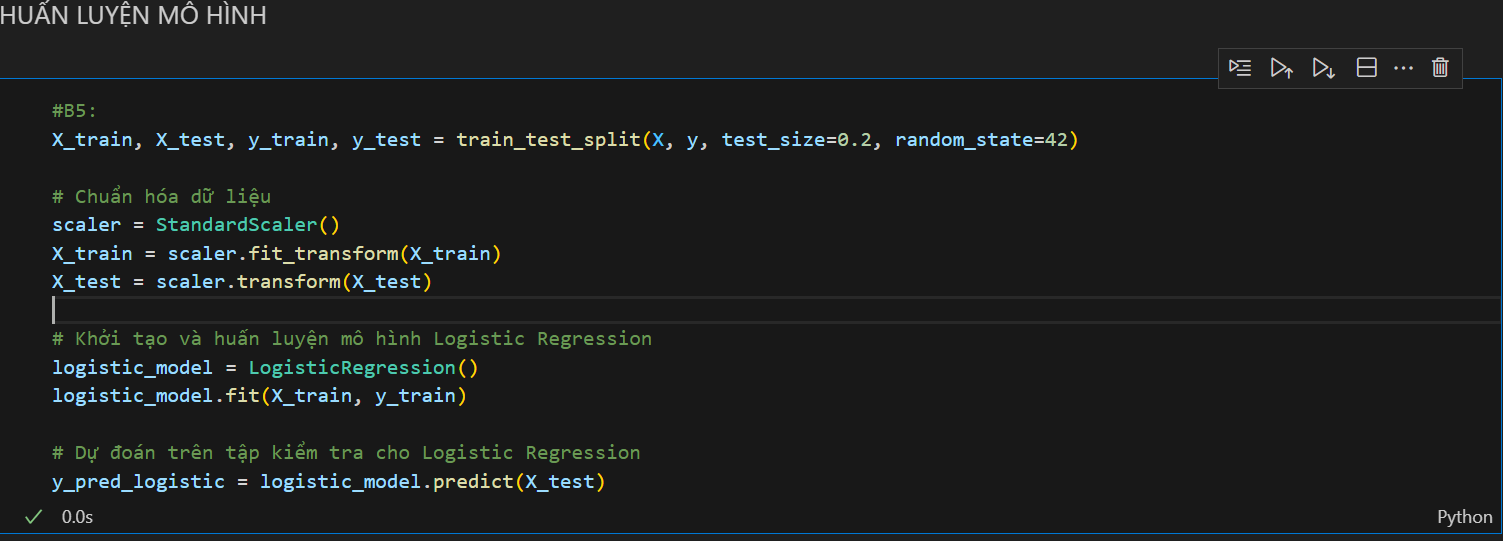
B3: Xác định các biến đầu vào và nhãn



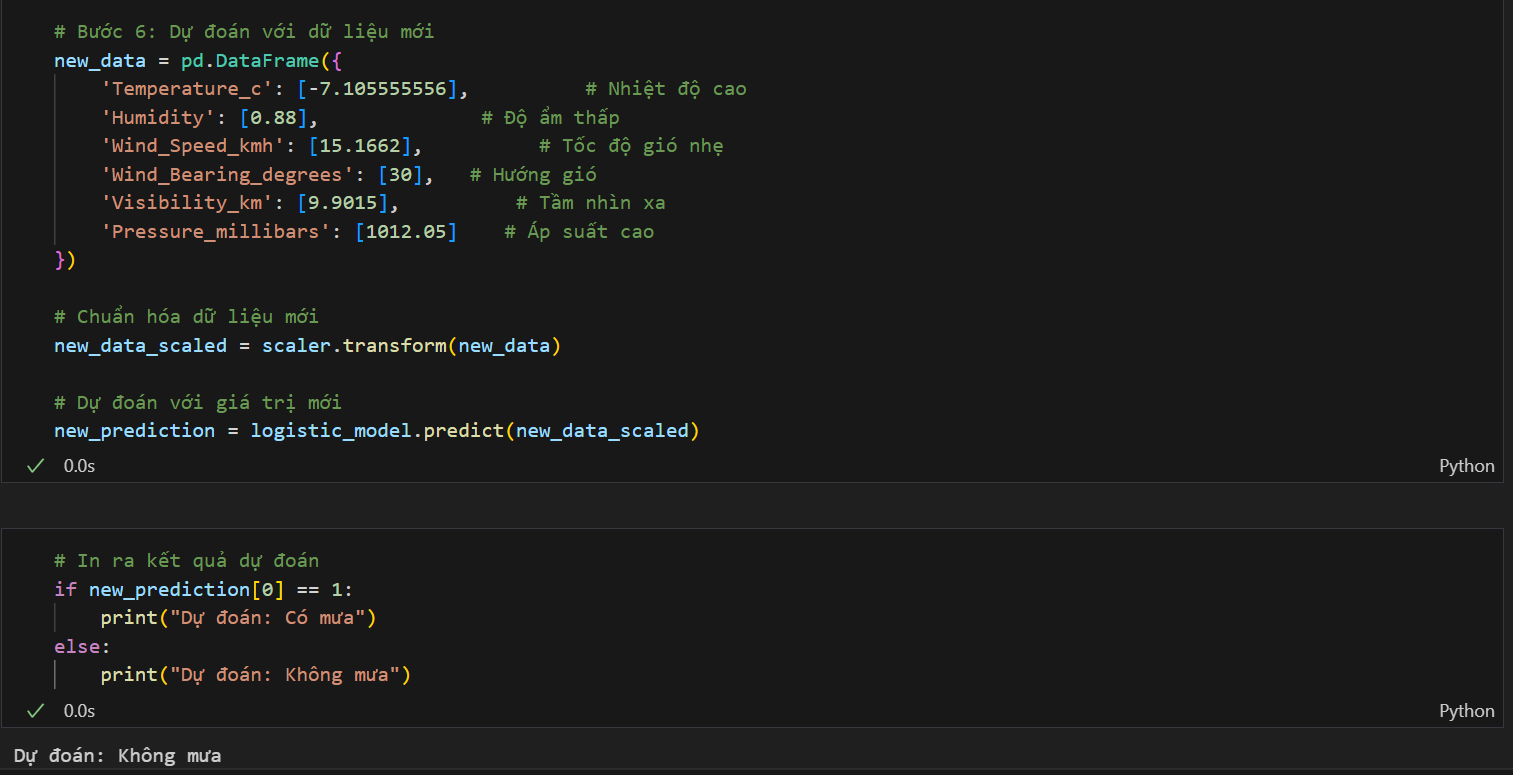
B4: Trực quan hóa dữ liệu bằng biểu đồ



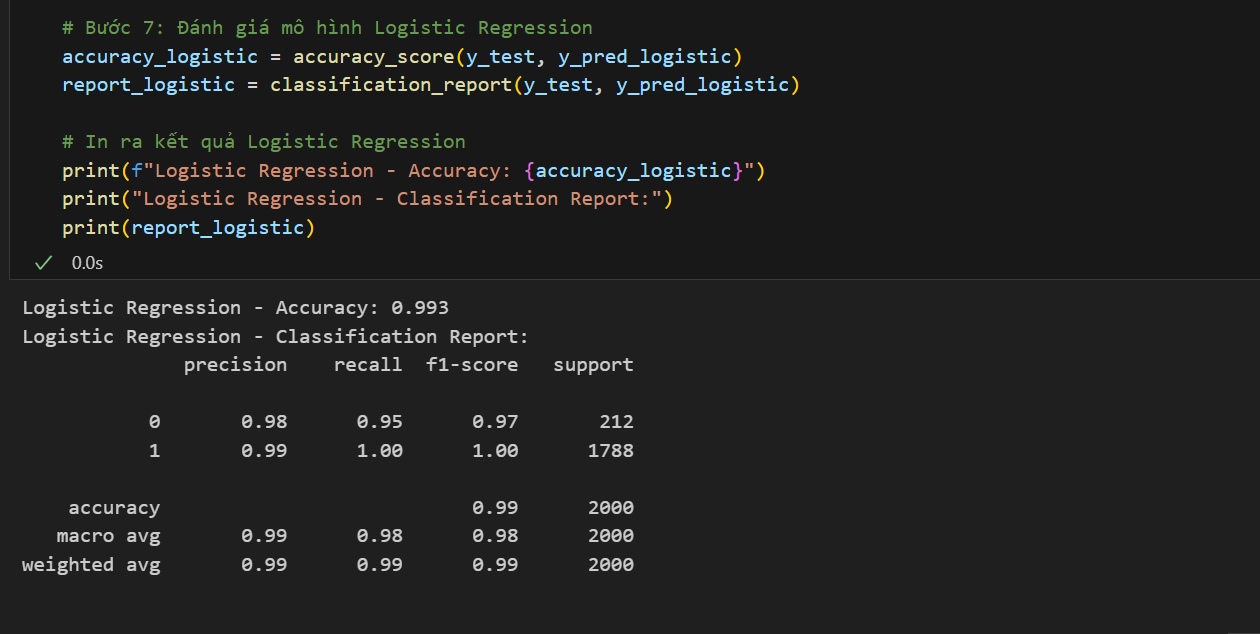
B5: Huấn luyện mô hình



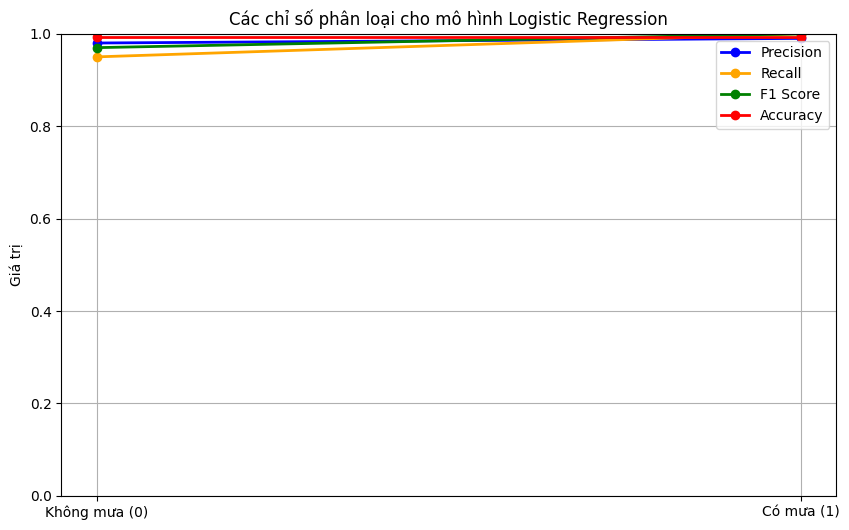
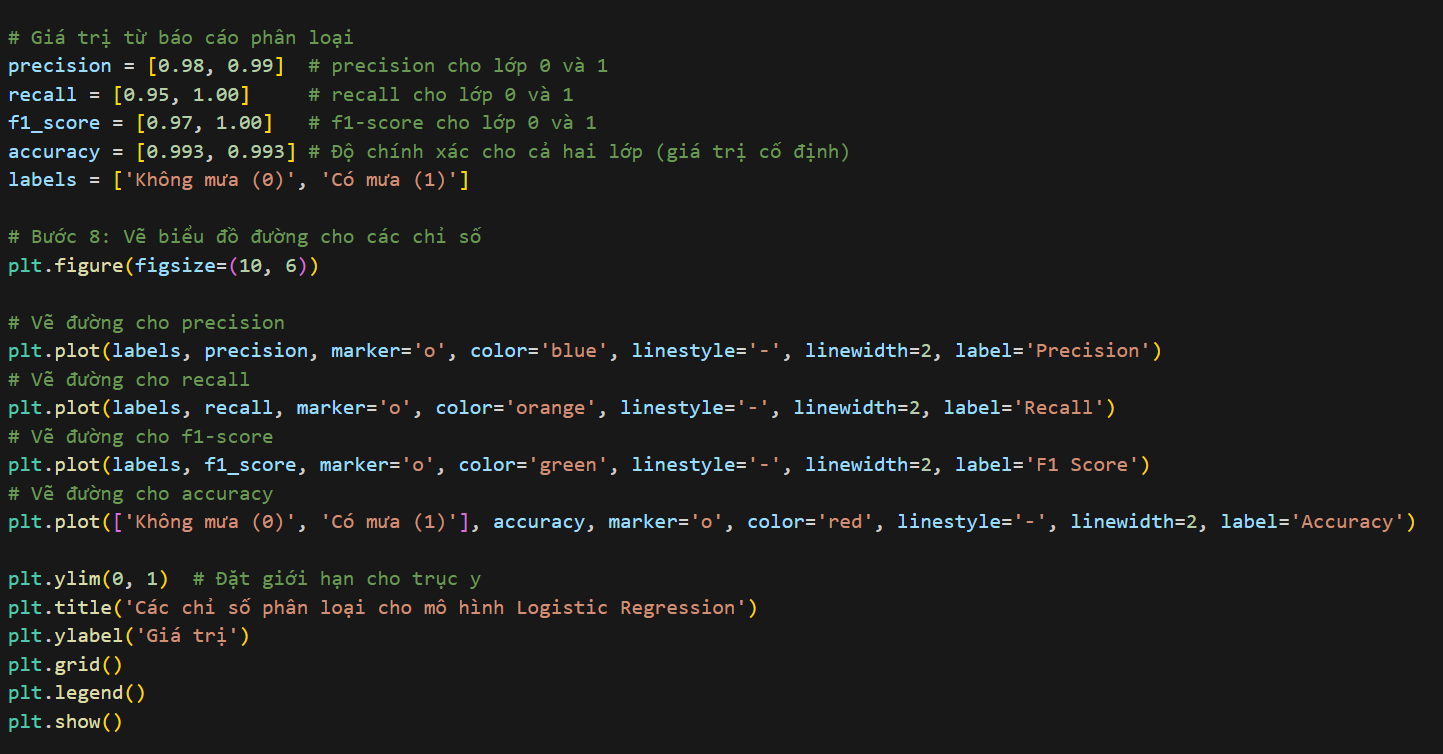
B6: Dự đoán dữ liệu mới



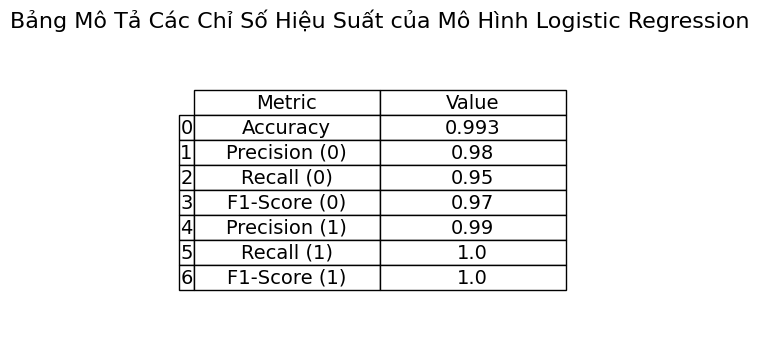
B7: Đánh giá mô hình



B8: Vẽ biểu đồ đường biểu diễn các chỉ số hiệu suất mô hình



**4. Phân tích kết quả của chương trình**



Từ bảng trên ta thấy được:

\* **Độ Chính Xác (Accuracy)**

* **Accuracy**: 0.993 (99.3%)

+ Điều này có nghĩa là 99.3% tổng số dự đoán của mô hình là chính xác. Độ chính xác này cho thấy mô hình hoạt động rất tốt trong việc phân loại dữ liệu.

\* **Precision, Recall, F1-Score**

* **Lớp 0 (Không có mưa)**

- **Precision**: 0.98

+ Nghĩa là trong số những lần mà mô hình dự đoán có 0 (không có mưa), 98% là chính xác.

- **Recall**: 0.95

+ Có nghĩa là trong số các trường hợp thực sự có 0 (không có mưa), mô hình đã phát hiện được 95% số trường hợp.

- **F1-Score**: 0.97

+ F1-Score = 0.97 biểu thị cho sự cân bằng tốt giữa Precision và Recall.

* **Lớp 1 (Có mưa)**

- **Precision**: 0.99

+ Nghĩa là trong số những lần mô hình dự đoán có 1 (có mưa), 99% là chính xác.

- **Recall**: 1.00

+ Có nghĩa là mô hình đã phát hiện hoàn toàn các trường hợp thực sự có 1 (có mưa).

- **F1-Score**: 1.00

+ Với F1-Score đạt 1.00, điều này cho thấy mô hình hoàn toàn chính xác trong việc phát hiện các trường hợp có mưa.

\* **Nhận Xét Cuối Cùng**

- Mô hình Logistic Regression thể hiện hiệu suất ấn tượng với độ chính xác đạt 99.3%, cho thấy khả năng phân loại tốt trong việc dự đoán có mưa hay không. Các chỉ số đánh giá như Precision và Recall đều ở mức cao, với Precision cho lớp 1 (có mưa) đạt 0.99 và Recall đạt 1.00, cho thấy mô hình có thể phát hiện hầu hết các trường hợp có mưa mà không bỏ lỡ. F1-Score đạt 0.97 cho lớp 0 (không mưa) và 1.00 cho lớp 1 cho thấy sự cân bằng tốt giữa độ chính xác và tỷ lệ phát hiện.

#### **Phần 3. Kết luận**

* Áp dụng được mô hình dự đoán thuật toán Logistic Regression
* Tìm ra tỷ lệ dự đoán đúng của thuật toán
* Đưa ra các thông tin kết quả mưa hay không mưa một cách tốt nhất

##### Tài liệu tham khảo

* Bài giảng của giảng viên: Tạ Quang Chiểu
* Một số tài liệu thêm: [Logistic Regression](https://machinelearningcoban.com/2017/01/27/logisticregression/)
* Data: [Rainfall](https://www.kaggle.com/datasets/zakriarehman/weather-data-for-linear-regression/data)