****

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**

A picture containing circle, logo, font, trademark

Description automatically generated

**ĐỀ TÀI: Phân tích dữ liệu thời tiết**

**Giảng viên: Phan Thành Huấn**

Tên sinh viên:

Nguyễn Ngọc Ánh - 3120410041

*Thành phố Hồ Chí Minh - Tháng 12/2023*

# LỜI CẢM ƠN

Chúng em xin chân thành gửi lời cảm ơn chân thành và sâu sắc nhất đến tập thể quý thầy cô trong khoa Công nghệ thông tin trường Đại học Sài Gòn đã giúp cho chúng em có được những kiến thức cơ bản làm nền tảng để hoàn tất thực hiện bài báo cáo này. Đặc biệt hơn, chúng em xin gửi lời cảm ơn và lòng biết ơn chân thành sâu sắc nhất tới thầy Phan Thành Huấn, giảng viên môn Phân tích dữ liệu. Thầy đã dành thời gian, nỗ lực để truyền đạt kiến thức, hướng dẫn tận tình và là người đồng hành trong quá trình học tập của chúng em. Thầy đã đóng góp nhiều ý kiến quý báu giúp cho chúng em hoàn thành tốt bài báo cáo môn học của mình.

Tuy nhiên, trong quá trình học tập và thực hiện, chúng em không tránh khỏi những thiếu sót và sai lầm. Chính vì thế, chúng em rất mong muốn và sẵn sàng lắng nghe được những sự góp ý từ phía thầy, nhằm hoàn thiện những kiến thức để nhóm của chúng em có thể học tập thêm và đó cũng là hành trang tốt nhất mà chúng em thực hiện tiếp các đề tài khác trong tương lai.

Một lần nữa, chúng em xin trân thành cảm thầy.

**MỤC LỤC**

[**LỜI CẢM ƠN 2**](#_gjdgxs)

[**Chương 1: TỔNG QUAN 4**](#_3qldkktnoxyt)

[I. Giới thiệu bài toán 4](#_r4d9wngpnq37)

[II. Hướng tiếp cận bài toán 4](#_ccdoqca0a9eu)

[**Chương 2: CƠ SỞ LÝ THUYẾT 5**](#_79ru90g2oetv)

[I. Tổng quan về học máy và phân tích dữ liệu 5](#_2s8eyo1)

[II. Tổng quan về học có giám sát và bài toán phân loại 6](#_3rdcrjn)

[III. Cây quyết định 7](#_lnxbz9)

[IV. Cơ sở lý thuyết về thời tiết 8](#_vff5w9amz9uc)

[V. Thuật toán Hồi quy Logistic 9](#_5ldr25q6bho7)

[**Chương 3: QUÁ TRÌNH THỰC NGHIỆM VÀ KẾT QUẢ 12**](#_pgdqalshaeou)

[1. Tổng quan về bộ dữ liệu 12](#_e8e57j3izjyc)

[2. Tiền xử lý dữ liệu 13](#_xga67lzbk2vx)

[3. Phân tích và đánh giá mối quan hệ giữa các dấu hiệu dự báo thời tiết 13](#_wfahvyizaxmo)

[4. Cây quyết định 18](#_x8y0o0g6nsij)

[**KẾT LUẬN – HƯỚNG PHÁT TRIỂN 21**](#_3l18frh)

[1. Kết quả đạt được 21](#_206ipza)

[2. Hạn chế của đồ án 21](#_4k668n3)

[3. Hướng phát triển 21](#_2zbgiuw)

# Chương 1: TỔNG QUAN

## Giới thiệu bài toán

Phân tích dữ liệu thời tiết để tìm hiểu mối quan hệ giữa các yếu tố như nhiệt độ, độ ẩm và mưa là một chủ đề nghiên cứu quan trọng trong lĩnh vực khoa học dữ liệu - giúp các nhà nghiên cứu hiểu rõ hơn về cách các yếu tố thời tiết khác nhau ảnh hưởng lẫn nhau, từ đó có thể dự báo thời tiết chính xác hơn và phát triển các giải pháp ứng phó với biến đổi khí hậu. Dữ liệu thời tiết được thu thập từ nhiều nguồn khác nhau, bao gồm các trạm khí tượng, vệ tinh và các cảm biến khác. Dữ liệu này thường bao gồm các thông tin như nhiệt độ, độ ẩm, áp suất, gió, lượng mưa, v.v. Phân tích dữ liệu thời tiết có thể được sử dụng để tìm hiểu mối quan hệ giữa các yếu tố thời tiết khác nhau. Ví dụ, các nhà nghiên cứu có thể sử dụng phân tích dữ liệu để tìm hiểu mối quan hệ giữa nhiệt độ và lượng mưa, hoặc giữa độ ẩm và áp suất. Kết quả của phân tích dữ liệu thời tiết có thể được sử dụng để cải thiện dự báo thời tiết. Dự báo thời tiết chính xác hơn có thể giúp các cá nhân và tổ chức chuẩn bị tốt hơn cho các sự kiện thời tiết khắc nghiệt, chẳng hạn như bão, lũ lụt và hạn hán. Phân tích dữ liệu thời tiết cũng có thể được sử dụng để nghiên cứu biến đổi khí hậu. Biến đổi khí hậu đang tác động đến thời tiết trên toàn thế giới. Phân tích dữ liệu thời tiết có thể giúp các nhà khoa học hiểu rõ hơn về cách biến đổi khí hậu đang thay đổi thời tiết và những tác động của nó đối với môi trường và con người.

## Hướng tiếp cận bài toán

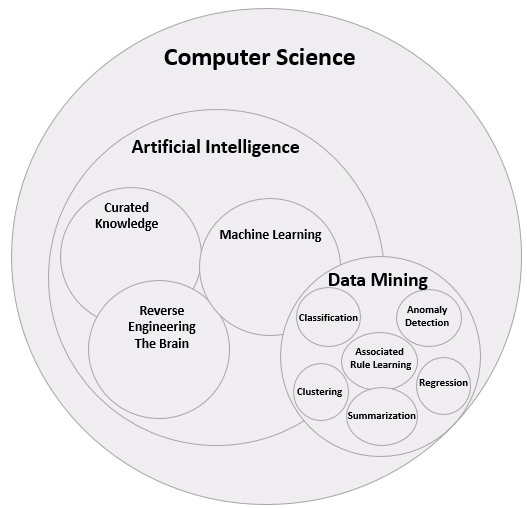
Phân tích dựa trên lý thuyết có sẵn: các nhà nghiên cứu sử dụng các lý thuyết khí tượng đã được biết đến để phân tích dữ liệu thời tiết. Các lý thuyết này có thể cung cấp những hiểu biết cơ bản về mối quan hệ giữa các yếu tố thời tiết khác nhau. Ví dụ, theo lý thuyết khí động học, nhiệt độ và áp suất có mối quan hệ với nhau. Khi nhiệt độ tăng, áp suất sẽ giảm. Các nhà nghiên cứu có thể sử dụng lý thuyết này để phân tích dữ liệu thời tiết để tìm hiểu mối quan hệ giữa nhiệt độ và áp suất.

Đào sâu thêm có mối quan hệ nào khác hay không: Sau khi đã có những hiểu biết cơ bản về mối quan hệ giữa các yếu tố thời tiết dựa trên lý thuyết, các nhà nghiên cứu sẽ tiến hành đào sâu thêm để tìm hiểu xem liệu có những mối quan hệ nào khác chưa được biết đến hay không. Các nhà nghiên cứu có thể sử dụng các kỹ thuật phân tích dữ liệu thống kê để tìm kiếm những mối quan hệ tiềm ẩn giữa các yếu tố thời tiết. Các kỹ thuật này có thể giúp các nhà nghiên cứu xác định những mối quan hệ có ý nghĩa thống kê, tức là những mối quan hệ không chỉ là ngẫu nhiên.

Chọn ra các cặp yếu tố: Từ những mối quan hệ đã được tìm thấy, các nhà nghiên cứu sẽ chọn ra các cặp yếu tố có mối quan hệ chặt chẽ nhất. Các cặp yếu tố này sẽ được sử dụng để xây dựng các thuật toán phân loại. Ví dụ, nếu các nhà nghiên cứu tìm thấy mối quan hệ chặt chẽ giữa nhiệt độ và lượng mưa, họ có thể sử dụng mối quan hệ này để xây dựng một thuật toán phân loại có thể dự đoán khả năng xảy ra mưa dựa trên nhiệt độ. Các thuật toán phân loại có thể được sử dụng để dự báo thời tiết, ứng phó với biến đổi khí hậu hoặc quản lý tài nguyên thiên nhiên. Hướng tiếp cận này giúp các nhà nghiên cứu hiểu rõ hơn về mối quan hệ giữa các yếu tố thời tiết, từ đó có thể sử dụng các thông tin này để giải quyết các vấn đề thực tế.

# Chương 2: CƠ SỞ LÝ THUYẾT

## Tổng quan về học máy và phân tích dữ liệu



Hình ảnh 3 Học máy và khai phá dữ liệu

Học máy (Machine Learning) và khai phá dữ liệu (Data Mining) là hai lĩnh vực quan trọng trong lĩnh vực trí tuệ nhân tạo và phân tích dữ liệu.

Học máy là một phương pháp của trí tuệ nhân tạo mà máy tính được huấn luyện để tự động học từ dữ liệu và cải thiện hiệu suất của nó trong một nhiệm vụ cụ thể mà không cần được lập trình một cách tường minh. Học máy giúp máy tính nhận biết các mẫu và quy luật trong dữ liệu và tạo ra các mô hình dự đoán hoặc phân loại dựa trên những mẫu đó. Các thuật toán Học máy bao gồm học có giám sát (supervised learning), học không giám sát (unsupervised learning) và học tăng cường (reinforcement learning). Học máy được áp dụng rộng rãi trong nhiều lĩnh vực như phân loại dữ liệu, nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên và nhiều ứng dụng khác.

Phân tích dữ liệu là quá trình thu thập, xử lý và phân tích dữ liệu để tìm ra thông tin có ý nghĩa. Phân tích dữ liệu có thể được sử dụng để giải quyết nhiều vấn đề khác nhau, từ việc cải thiện hiệu quả hoạt động của doanh nghiệp đến việc phát triển các sản phẩm và dịch vụ mới. Phân tích dữ liệu thường được chia thành các bước sau:

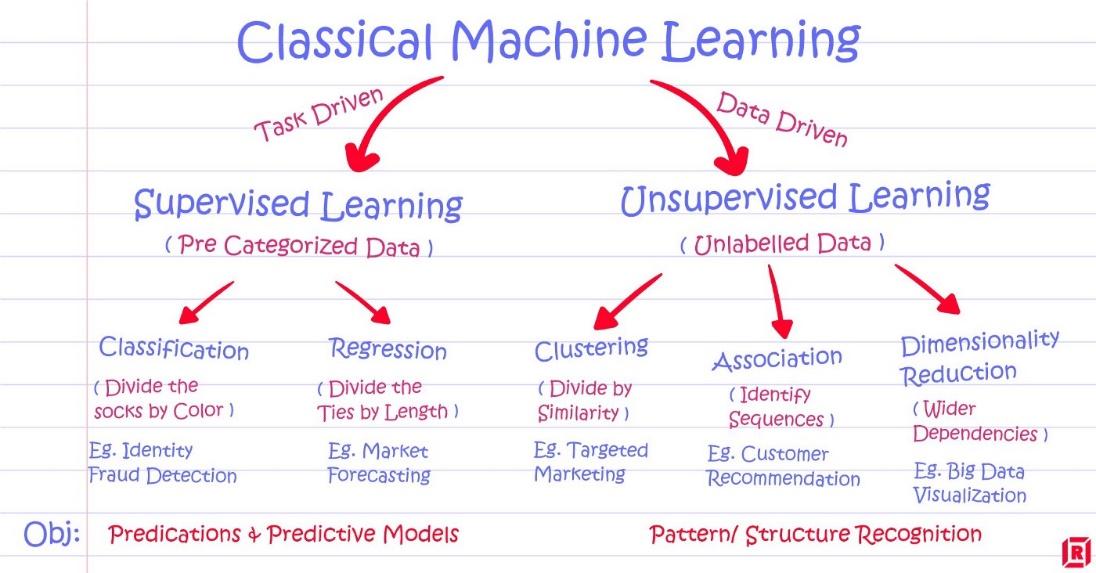
* Thu thập dữ liệu: Bước đầu tiên là thu thập dữ liệu từ nhiều nguồn khác nhau, chẳng hạn như các bảng tính, cơ sở dữ liệu, trang web hoặc các thiết bị cảm biến;
* Xử lý dữ liệu: Sau khi thu thập dữ liệu, cần phải xử lý dữ liệu để loại bỏ các lỗi và khiếm khuyết;
* Phân tích dữ liệu: Bước này sử dụng các kỹ thuật thống kê và khai thác dữ liệu để tìm hiểu ý nghĩa của dữ liệu;
* Trực quan hóa dữ liệu: Bước này sử dụng các biểu đồ và đồ thị để giúp người dùng dễ dàng hiểu các kết quả phân tích dữ liệu.

Sự kết hợp của Học máy và phân tích dữ liệu cung cấp công cụ mạnh mẽ để xử lý và trích xuất tri thức từ dữ liệu, từ đó mang lại lợi ích và giá trị cho các tổ chức và cá nhân.

## Tổng quan về học có giám sát và bài toán phân loại

**Học có giám sát (supervised learning)** là một phương pháp trong lĩnh vực bài học máy mà mô hình được huấn luyện từ các cặp dữ liệu đầu vào và đầu ra tương ứng. Trong quá trình huấn luyện, mô hình học các mẫu từ dữ liệu đã được gán nhãn để tạo ra một mô hình dự đoán có khả năng dự đoán đầu ra cho các dữ liệu mới chưa được gán nhãn. Quá trình học có giám sát bắt đầu bằng việc chuẩn bị dữ liệu huấn luyện, trong đó mỗi mẫu dữ liệu được gắn kết với một nhãn đúng. Nhãn này có thể là một giá trị số (ví dụ: dự đoán giá tiền nhà dựa trên diện tích) hoặc một nhãn lớp (ví dụ: phân loại email vào hộp thư rác hoặc hộp thư đến). Sau đó, mô hình được huấn luyện bằng cách tìm cách tối ưu hóa hàm mất mát, mục tiêu là giảm thiểu sai số giữa đầu ra dự đoán và nhãn đúng của dữ liệu huấn luyện. Các thuật toán phổ biến trong học có giám sát bao gồm hồi quy tuyến tính, cây quyết định, máy vector hỗ trợ (SVM), mạng neural, và rất nhiều thuật toán khác. Sau khi mô hình đã được huấn luyện, nó có thể được sử dụng để dự đoán đầu ra cho các dữ liệu mới chưa được gán nhãn. Quá trình này được gọi là kiểm tra hoặc đánh giá mô hình, và độ chính xác của dự đoán được đánh giá bằng các phương pháp như ma trận nhầm lẫn, độ chính xác, độ nhạy, độ đặc hiệu, và các độ đo khác. Học có giám sát có ứng dụng rộng rãi trong nhiều lĩnh vực, bao gồm phân loại ảnh, dịch thuật tự động, nhận dạng giọng nói, hệ thống khuyến nghị, và nhiều ứng dụng khác trong việc dự đoán và phân loại dữ liệu.

**Bài toán phân loại** là một trong những bài toán quan trọng trong lĩnh vực máy học và khai phá dữ liệu. Nhiệm vụ của bài toán phân loại là xây dựng một mô hình có khả năng phân loại các điểm dữ liệu vào các nhóm, lớp hoặc danh mục khác nhau dựa trên các đặc trưng (features) của chúng. Quá trình phân loại thường bắt đầu bằng việc chuẩn bị dữ liệu huấn luyện, trong đó mỗi điểm dữ liệu được gán nhãn đúng thuộc về nhóm nào. Dữ liệu huấn luyện này được sử dụng để huấn luyện mô hình phân loại, với mục tiêu là tìm ra một quy tắc, mô hình hoặc hàm quyết định để phân loại chính xác các điểm dữ liệu mới. Có nhiều phương pháp và thuật toán phân loại khác nhau, bao gồm cây quyết định, hồi quy logistic, máy vector hỗ trợ (SVM), mạng neural, k-Nearest Neighbors (k-NN), Random Forest, và Gradient Boosting, chỉ định cùng một số. Mỗi thuật toán có cách hoạt động và ưu điểm riêng, và lựa chọn thuật toán phù hợp phụ thuộc vào loại dữ liệu và yêu cầu của bài toán. Sau khi mô hình đã được huấn luyện, nó có thể được sử dụng để phân loại các điểm dữ liệu mới, tức là dự đoán nhãn của chúng. Độ chính xác của mô hình được đánh giá thông qua các phương pháp như ma trận nhầm lẫn, độ chính xác, độ nhạy, độ đặc hiệu, và các độ đo khác. Bài toán phân loại có ứng dụng rộng rãi trong nhiều lĩnh vực, bao gồm phân loại email, phát hiện gian lận tín dụng, nhận dạng hình ảnh, phân loại văn bản, và nhiều ứng dụng khác trong việc tự động phân loại và quyết định dựa trên dữ liệu.



Hình ảnh 4 Học có giám sát và bài toán phân loại

## Cây quyết định

Cây quyết định là một thuật toán trong lĩnh vực học máy. Nó được sử dụng để tạo ra mô hình dự đoán dựa trên việc học từ dữ liệu đào tạo. Cây quyết định có thể được sử dụng cho cả bài toán phân loại và dự đoán. Một cây quyết định bao gồm một cây có cấu trúc phân nhánh, trong đó mỗi nút trong cây đại diện cho một thuộc tính và mỗi nhánh của nút đại diện cho một giá trị của thuộc tính đó. Tại mỗi nút lá của cây, ta tìm được một kết quả hoặc một dự đoán.

Quá trình xây dựng cây quyết định bắt đầu từ nút gốc của cây và tiếp tục phân tách các nút con dựa trên thuộc tính tốt nhất để phân chia dữ liệu. Các thuộc tính được đánh giá dựa trên các phép đo, chẳng hạn như độ tinh khiết (gini impurity) hoặc thông tin độ (information gain). Mục tiêu là tạo ra một cây có khả năng phân loại tốt nhất và dự đoán chính xác nhất trên dữ liệu mới.

Cây quyết định có một số lợi ích như dễ hiểu và diễn giải, có khả năng làm việc với các biến định tính và số học, và có khả năng xử lý các dữ liệu thiếu. Tuy nhiên, nó cũng có một số hạn chế, bao gồm khả năng quá khớp (overfitting) dữ liệu đào tạo và nhạy cảm với các thay đổi nhỏ trong dữ liệu đầu vào.

Để tránh quá khớp và cải thiện hiệu suất của cây quyết định, có một số biến thể và phương pháp kết hợp, bao gồm cắt tỉa (pruning), sử dụng cây ngẫu nhiên (random forests) và bổ sung cây (boosting). Những phương pháp này giúp cải thiện tính tổng quát và ổn định của cây quyết định.

## Cơ sở lý thuyết về thời tiết

Thời tiết là các hiện tượng tự nhiên được con người quan sát trong một thời gian dài. Thông qua quá trình quan sát, con người đã rút ra được các yếu tố là nguyên nhân dẫn đến sự thay đổi của thời tiết.

Trong bài toán dự đoán trời mưa, tôi chỉ tập trung vào các yếu tố có thể ảnh hưởng đến trời mưa hoặc lượng mưa tại một khu vực. Mưa xuất hiện do quá trình ngưng tụ hơi nước trong không khí, sau đó các hạt nước tụ lại để tạo thành giọt nước đủ lớn để rơi xuống mặt đất và trở thành mưa. Lượng mưa là số lượng nước mưa rơi xuống một khu vực cụ thể trong một khoảng thời gian nhất định, thường được đo bằng đơn vị đo lường thích hợp như milimet (mm) hoặc inch. Lượng mưa quan trọng trong việc đánh giá khí hậu, thực hiện quản lý tài nguyên nước và nông nghiệp, cũng như dự đoán và ứng phó với các sự kiện thời tiết đối với môi trường và xã hội.

Những nhân tố có thể ảnh hưởng đến lượng mưa là các yếu tố tự nhiên và địa lý có vai trò quan trọng trong việc xác định mức độ và phân bố lượng mưa trên một khu vực cụ thể. Sau đây là một số các yếu tố:

* Khí áp: Trong khu vực áp suất thấp, không khí ẩm và nóng được nâng lên và ngưng tụ thành mưa. Khí áp thất có thể gây ra sự tăng cường của quá trình ngưng mưa. Trong khu vực khí áp suất cao, không khí thoát ra và làm cho không khí xung quanh trở nên khô ráo.
* Gió: Những vùng nằm sâu trong các lực địa nếu không có gió từ đại dương thổi vào thì mưa ít. Mưa ở các khu vực này thường do ngưng kết hơi nước từ ao hồ, sông và rừng cây bốc lên. Ngoài ra, Miền có gió Mậu dịch ít mưa do gió này chủ yếu là gió khô, trong khi đó, miền có gió mùa mưa nhiều vì gió mùa mùa hạ thổi từ đại dương vào đem theo nhiều hơi nước.
* Dòng biển: Bờ đại dương gần nơi có dòng biển nóng đi qua thường có mưa nhiều vì không khí trên dòng biển nóng chứa nhiều hơi nước. Ngược lại, bờ đại dương có dòng biển lạnh đi qua mưa ít vì không khí trên dòng biển bị lạnh, hơi nước không bốc lên được.

Trong phần thực nghiệm dưới, tôi sẽ có sử dụng các yếu tố khác ngoài những yếu tố tự nhiên trên để tìm mối quan hệ giữa các yếu tố và mối quan hệ với lượng mưa.

## Thuật toán Hồi quy Logistic

Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

Hồi quy logistic là một kỹ thuật quan trọng trong lĩnh vực trí tuệ nhân tạo và máy học (AI/ML). Mô hình ML là các chương trình phần mềm có thể được đào tạo để thực hiện các tác vụ xử lý dữ liệu phức tạp mà không cần sự can thiệp của con người. Mô hình ML được xây dựng bằng hồi quy logistic có thể giúp các tổ chức thu được thông tin chuyên sâu hữu ích từ dữ liệu kinh doanh của mình. Họ có thể sử dụng những thông tin chuyên sâu này để phân tích dự đoán nhằm giảm chi phí hoạt động, tăng độ hiệu quả và đổi chỉnh quy mô nhanh hơn. Ví dụ: doanh nghiệp có thể khám phá các mẫu hình cải thiện khả năng giữ chân nhân viên hoặc tạo ra thiết kế sản phẩm mang về nhiều lợi nhuận hơn.

Hồi quy logistic là một trong số ít các kỹ thuật phân tích hồi quy khác nhau thường được các nhà khoa học dữ liệu sử dụng trong máy học (ML). Để hiểu rõ về hồi quy logistic, trước tiên chúng ta phải hiểu kỹ thuật phân tích hồi quy cơ bản. Dưới đây là một ví dụ về phân tích hồi quy tuyến tính để cho thấy cách thức hoạt động của phân tích hồi quy.

**Xác định câu hỏi**

Bất kỳ quá trình phân tích dữ liệu nào cũng bắt đầu bằng một câu hỏi kinh doanh. Đối với hồi quy logistic, bạn nên giới hạn phạm vi câu hỏi để có được kết quả cụ thể:

* Những ngày mưa có ảnh hưởng đến doanh số hàng tháng của chúng ta không? (có hoặc không)
* Khách hàng đang thực hiện loại hoạt động thẻ tín dụng nào? (ủy quyền, gian lận hoặc có khả năng gian lận)

**Thu thập dữ liệu lịch sử**

Sau khi xác định câu hỏi, bạn cần xác định các yếu tố dữ liệu có liên quan. Sau đó, bạn sẽ thu thập dữ liệu trước đây cho tất cả các yếu tố. Ví dụ: để trả lời câu hỏi đầu tiên ở trên, bạn có thể thu thập dữ liệu doanh số hàng tháng và số ngày mưa mỗi tháng trong ba năm qua.

**Đào tạo mô hình phân tích hồi quy**

Bạn sẽ xử lý dữ liệu lịch sử bằng phần mềm hồi quy. Phần mềm sẽ xử lý các điểm dữ liệu khác nhau và kết nối chúng theo phương thức toán học bằng cách sử dụng phương trình. Ví dụ: nếu số ngày mưa trong ba tháng là 3, 5 và 8 còn doanh số trong những tháng đó là 8, 12 và 18, thuật toán hồi quy sẽ kết nối các yếu tố này với phương trình:

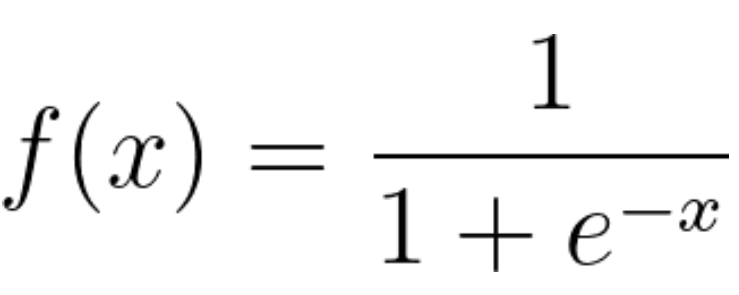
Doanh số = 2\*(Số ngày mưa) + 2

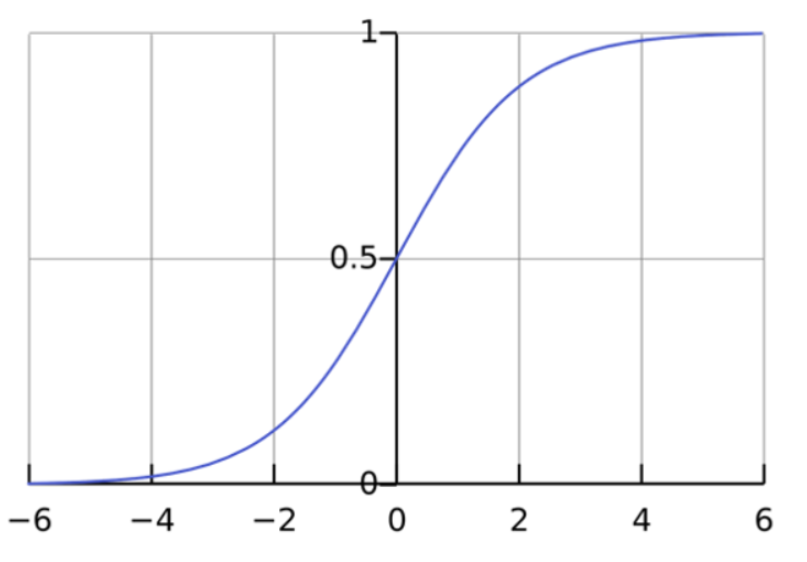
**Dự đoán các giá trị không xác định**

Đối với các giá trị không xác định, phần mềm sẽ áp dụng phương trình để dự đoán. Nếu bạn biết rằng tháng 7 sẽ có sáu ngày mưa, phần mềm sẽ ước tính giá trị doanh số trong tháng 7 là 14.

Hàm hồi quy logistic

Hồi quy logistic là một mô hình thống kê sử dụng hàm logistic, hay hàm logit trong toán học làm phương trình giữa x và y. Hàm logit ánh xạ y làm hàm sigmoid của x.

Hồi quy logistic sử dụng hàm logistic để mô tả mối quan hệ giữa biến phụ thuộc và biến độc lập. Hàm logistic là một hàm phi tuyến có thể chuyển đổi giá trị của biến độc lập thành một giá trị nằm trong khoảng từ 0 đến 1.

Nếu vẽ phương trình hồi quy logistic này, bạn sẽ có một đường cong hình chữ S như hình dưới đây.

Như bạn có thể thấy, hàm logit chỉ trả về các giá trị giữa 0 và 1 cho biến phụ thuộc, dù giá trị của biến độc lập là gì. Đây là cách hồi quy logistic ước tính giá trị của biến phụ thuộc. Phương pháp hồi quy logistic cũng lập mô hình phương trình giữa nhiều biến độc lập và một biến phụ thuộc.

# Chương 3: QUÁ TRÌNH THỰC NGHIỆM VÀ KẾT QUẢ

## Tổng quan về bộ dữ liệu

Tập dữ liệu: Weather Prediction

Nguồn: [https://www.kaggle.com/datasets/thedevastator/weather-prediction?fbclid=Iwa](https://www.kaggle.com/datasets/thedevastator/weather-prediction?fbclid=IwAR2izdvl82F2fySprmE3an3pw1lfF8BRhx9dfAYhN5NsGWdjpGDL212G0l8)

Dữ liệu khí tượng ban đầu được lấy từ EUROPEAN CLIMATE ASSESSMENT & DATASET, một dự án cung cấp các quan sát hàng ngày tại các trạm khí tượng trên khắp châu Âu và Địa Trung Hải, trong đó 18 thành phố hoặc địa điểm khác nhau ở châu Âu đã được lựa chọn, đó là Basel (Thụy Sĩ), Budapest (Hungary), Dresden, Düsseldorf, Kassel, München (tất cả đều ở Đức), De Bilt và Maastricht (Hà Lan), Heathrow (Anh), Ljubljana (Slovenia), Malmo và Stockholm (Thụy Điển), Montélimar, Perpignan và Tours (Pháp), Oslo (Na Uy), Roma (Ý) và Sonnblick (Áo).

Bản ghi các quan sát khí tượng hàng ngày cho 18 địa điểm này kéo dài qua các thời điểm khác nhau. Tuy nhiên, ở đây, chúng tôi chỉ chọn khoảng thời gian từ năm 2000 đến năm 2010, dẫn đến 3654 quan sát hàng ngày.

Bên cạnh các yếu tố thời tiết chung được ghi lại tại tất cả các địa điểm như ‘nhiệt độ trung bình’, ‘Nhiệt độ tối đa’, ‘Nhiệt độ tối thiểu’, tập dữ liệu cũng được thêm vào các yếu tố khác như ‘Độ che phủ của mây’, ‘Tốc độ gió’,...

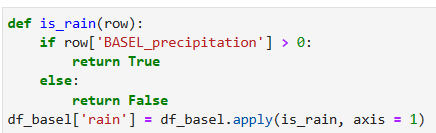
Sau khi thu thập dữ liệu, đã thực hiện dọn dẹp dữ liệu rất cơ bản. Các cột có > 5% mục không hợp lệ (“-9999”) đã được xóa, các cột có <= 5% mục không hợp lệ đã được giữ lại nhưng các mục không hợp lệ đã được thay thế bằng giá trị trung bình. Điều này dẫn đến 165 biến (hoặc tính năng) trong khoảng thời gian 3654 ngày.

Cuối cùng, tác giả đã chuyển đổi một số đơn vị dữ liệu để đạt được các phạm vi tương tự hơn của các giá trị hiện tại. Điều này làm cho dữ liệu phù hợp hơn để sử dụng cho học máy hoặc học sâu ngay cả khi không cần xử lý thêm. Tác giả cố tình không chọn chuẩn hóa hoàn toàn dữ liệu vì tác giả muốn giữ các đơn vị và giá trị được trình bày càng dễ tiếp cận một cách trực quan càng tốt. Nhiệt độ bây giờ được cho bằng độ Celsius, tốc độ gió và gió giật bằng m/s, độ ẩm bằng phần trăm của 100%, áp suất mực nước biển bằng 1000 hPa, bức xạ toàn cầu bằng 100 W/m2, lượng mưa bằng cm, nắng bằng giờ.

| Biến | Đơn vị gốc | Đơn vị đã chuyển đổi |
| --- | --- | --- |
| CC | Độ che phủ mây bằng okta | Độ che phủ mây bằng okta |
| DD | Hướng gió bằng độ | Hướng gió bằng độ |
| FG | Tốc độ gió bằng 0,1 m/s | Tốc độ gió bằng 1 m/s |
| FX | Gió giật bằng 0,1 m/s | Gió giật bằng 1 m/s |
| HU | Độ ẩm trong 1% | Độ ẩm trong tỷ lệ phần trăm của 100% |
| PP | Áp suất mực nước biển bằng 0,1 hPa | Áp suất mực nước biển bằng 1000 hPa |
| QQ | Bức xạ toàn cầu bằng W/m2 | Bức xạ toàn cầu bằng 100 W/m2 |
| RR | Lượng mưa bằng 0,1 mm | Lượng mưa bằng 10 mm |
| SS | Nắng trong 0,1 giờ | Nắng trong 1 giờ |
| TG | Nhiệt độ trung bình bằng 0,1 °C | Nhiệt độ trung bình bằng 1 °C |
| TN | Nhiệt độ tối thiểu bằng 0,1 °C | Nhiệt độ tối thiểu bằng 1 °C |
| TX | Nhiệt độ tối đa bằng 0,1 °C | Nhiệt độ tối đa bằng 1 °C |

## Tiền xử lý dữ liệu

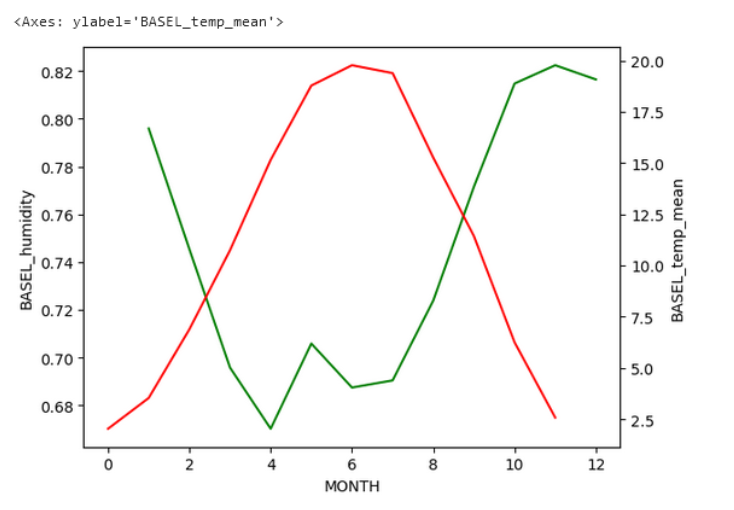
Trong dataset ban đầu, các bộ nhãn được sử dụng với mục đích của dataset là kiểm tra xem thời tiết vào ngày hôm đó, với tình hình có lượng mây, hướng gió, tốc độ của gió, độ ẩm, nhiệt độ như thế thì có thật sự thích hợp để đi picnic (dã ngoại) hay là không. Tuy nhiên, với mục tiêu đề tài là phân tích thời tiết, dự đoán vào ngày hôm đó với tình hình như vậy có thực sự xảy ra mưa hay là không? Vì vậy, em đã thay đổi dataset gốc bằng cách tạo thêm một cột dữ liệu tên là “Rain” và có kiểu dữ liệu là “Boolean”. Sau đó, lấy giá trị lượng mưa kiểm tra xem có lớn hơn 0 không? Nếu có thì trả về giá trị **True** hoặc ngược lại trả về giá trị **False.**



Dữ liệu được chia nhỏ thành từng thành phố để dễ dàng phân tích các mối quan hệ tại thành phố.

## Phân tích và đánh giá mối quan hệ giữa các dấu hiệu dự báo thời tiết

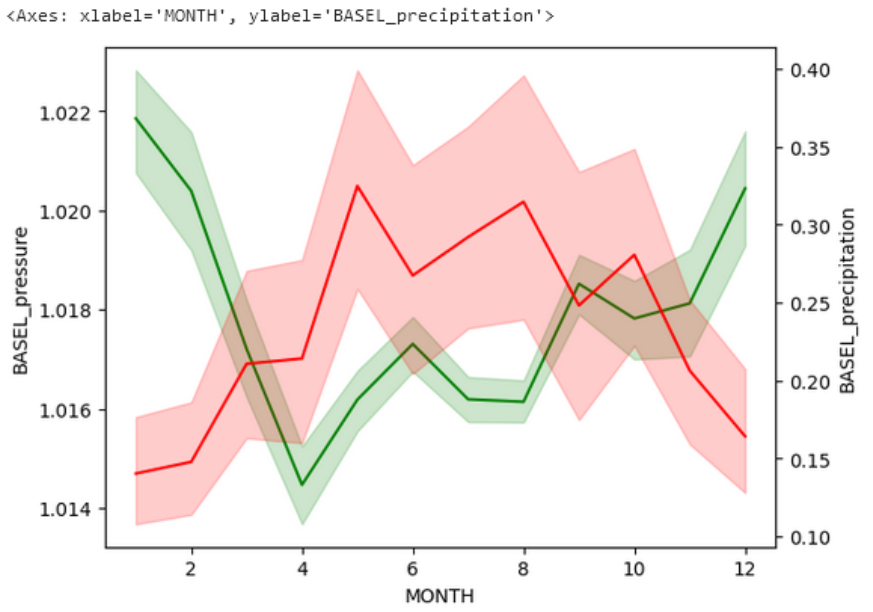
* Vẽ chart để thể hiện mối liên quan giữa độ ẩm và nhiệt độ



Nhìn vào chart đồ thị thì ta sẽ thấy rằng trong 1 năm nhiệt độ luôn có sự thay đổi theo các tháng và cao nhất vào tầm tháng 567 - thấp nhất tháng 11-1, độ ẩm cũng thay đổi theo mùa và cao nhất tầm 10 - 12 - thấp nhất tháng 4-7. Như vậy có sự chênh lệch rõ rệt giữa nhiệt độ và độ ẩm, nhiệt độ càng cao thì độ ẩm thấp và nhiệt độ thấp thì độ ẩm cao. Và độ ẩm cũng liên quan đến lượng mưa. Độ ẩm cao thì lượng mưa rất ít.

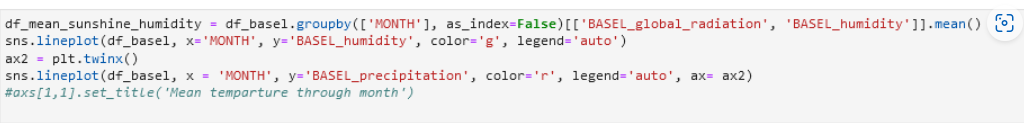
* Vẽ chart để thể hiện mối liên quan giữa lượng mưa và áp suất mực nước biển

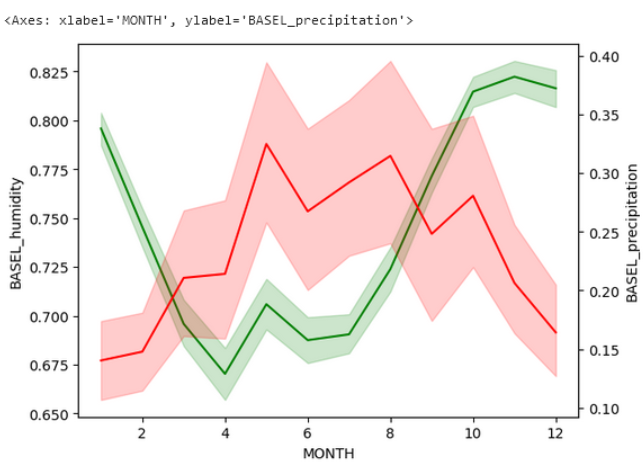




Từ chart, trong 1 năm 12 tháng đều có sự thay đổi về lượng mưa và áp suất mực nước biển. Lượng mưa nhiều vào tầm tháng 5-8 - thấp nhất tháng 12-1, trong khi đó áp suất cao nhất tháng 12-1, thấp nhất 3-5. Tuy sự chênh lệch giữa lượng mưa và áp suất không lớn nhưng áp suất cao thì lượng mưa thấp và ngược lại, áp suất thấp thì lượng mưa cao.

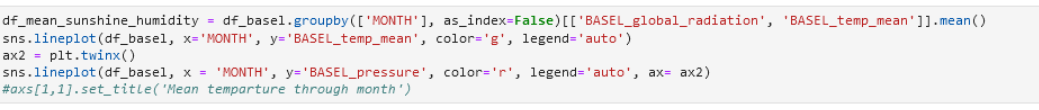
* Vẽ chart để thể hiện mối liên quan giữa lượng mưa và độ ẩm

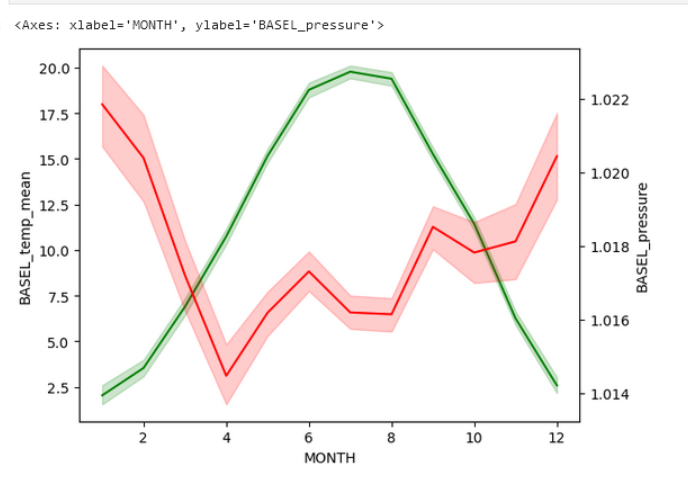




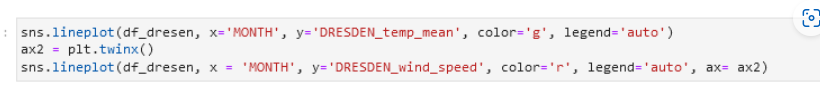
Từ chart, trong 1 năm 12 tháng đều có sự thay đổi về lượng mưa và độ ẩm. Lượng mưa nhiều vào tầm tháng 5-8 - thấp nhất tháng 12-1, trong khi đó độ ẩm cao nhất tháng 11-1, thấp nhất 3-5. Sự chênh lệch giữa lượng mưa và độ ẩm cho thấy rằng mưa nhiều thì độ ẩm không khí thấp và ngược lại, mưa ít thì độ ẩm không khí cao.

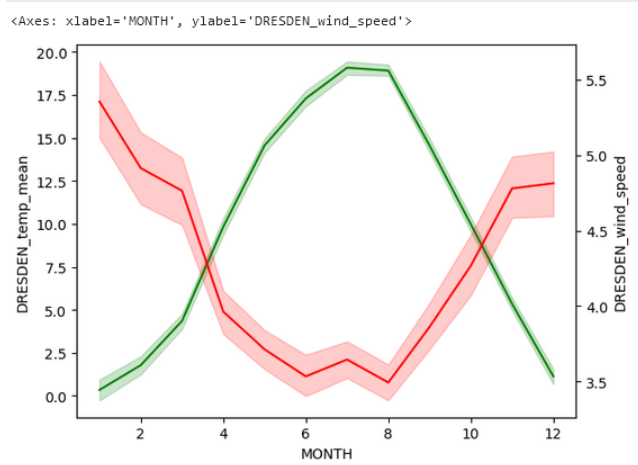
* Vẽ chart thể hiện mối liên quan giữa nhiệt độ và áp suất



Từ chart, trong 1 năm 12 tháng đều có sự thay đổi về nhiệt độ và áp suất. Trong 1 năm nhiệt độ luôn có sự thay đổi theo các tháng và cao nhất vào tầm tháng 567 - thấp nhất tháng 11-1, trong khi đó áp suất cao nhất tháng 12-1, thấp nhất 3-5. Như vậy, áp suất và nhiệt độ cũng chênh lệch trái ngược nhau, nhiệt độ cao thì áp suất thấp và ngược lại, nhiệt độ thấp đi thì áp suất tăng cao. Mà áp suất cao thì lượng mưa thấp vì vậy nhiệt độ thấp thì lượng mưa cũng thấp.

* Vẽ chart thể hiện mối liên quan giữa tốc độ gió và nhiệt độ



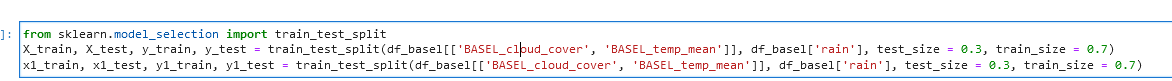


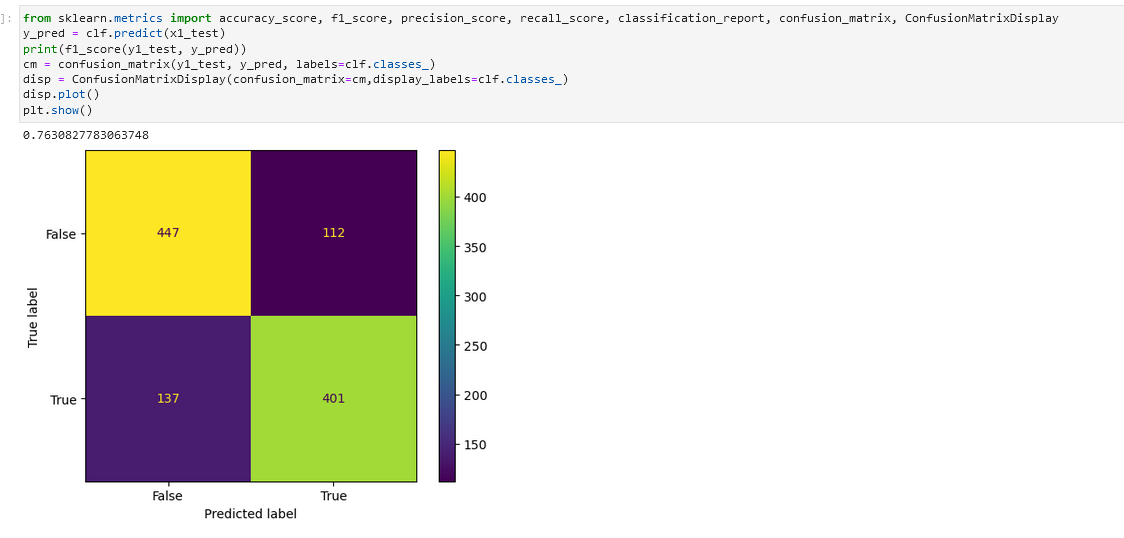
Khi qua thành phố khác, thì từ chart có thể thấy rằng, trong 1 năm, nhiệt độ luôn có sự thay đổi theo các tháng và cao nhất vào tầm tháng 567 - thấp nhất tháng 11-1, trong khi đó tốc độ gió mạnh tầm tháng 12-2 - yếu vào tháng 5-8. Nhìn sự chênh lệch của chart, tốc độ gió mạnh thì nhiệt độ hạ thấp và ngược lại, tốc độ gió yếu thì nhiệt độ tăng cao. Như vậy tốc độ gió cũng có ảnh hưởng đến lượng mưa

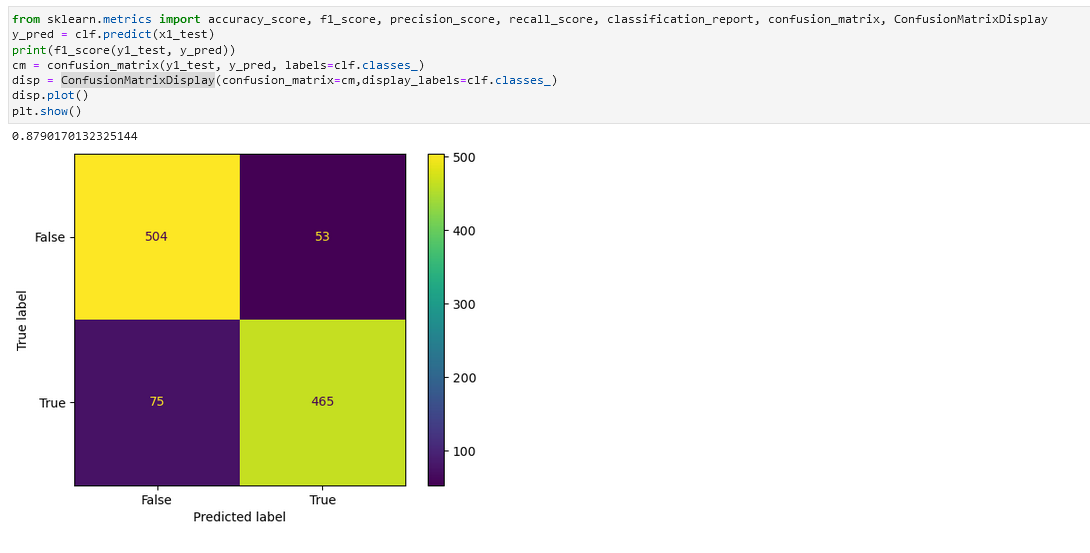
Có thể nhận thấy rằng nhiệt độ, độ ẩm, lượng mưa, áp suất, tốc độ gió luôn thay đổi và sự thay đổi của yếu tố này cũng làm thay đổi yếu tố còn lại.

## Cây quyết định

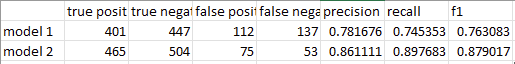
Trước khi huấn luyện, em chia tập dữ liệu thành 2 tập: tập huấn luyện (training set – 70% tập dữ liệu) và tập kiểm tra (test set – 30% tập dữ liệu).

Sau khi xác định nhiệt độ và độ che phủ của mây có ảnh hưởng đến lượng mưa, em sử dụng 2 yếu tố này dùng để huấn luyện và dự đoán lượng mưa.





Sau khi dùng mô hình f1-score, thì em thu nhận được kết quả như sau:

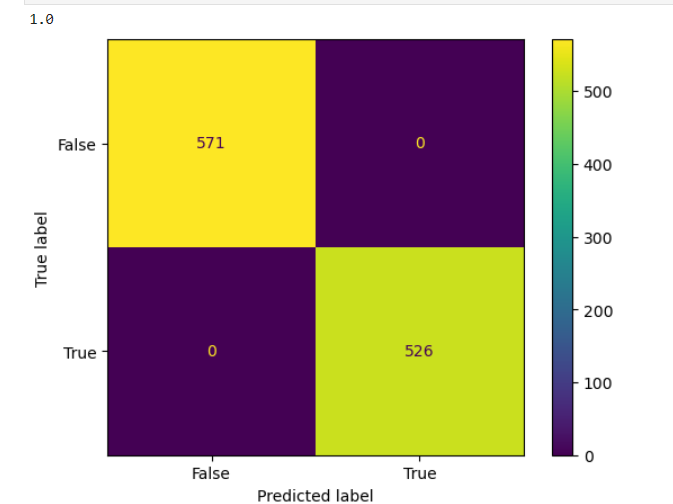


Các giá trị dự đoán với kết quả True-True và kết quả False-False có tỉ lệ là 848/1097 ≈ 0.77. Như vậy chỉ với 2 yếu tố nhiệt độ và độ che phủ thì có thể đưa ra kết quả trời có mưa hay là không với độ chính xác khoảng 77%. Nhưng trời mưa cũng còn nhiều yếu tố khác nữa.

Sau khi đổi qua thành phố khác với sự xuất hiện của yếu tố tốc độ gió và nhận thấy tốc độ gió có ảnh hưởng lượng mưa, em đã thêm tốc độ gió vào tham số huấn luyện DecisionTree, độ chính xác của mô hình tăng đáng kể. Tỉ lệ True-True và False-False trong lần thêm yếu tố là 969/1097 ≈ 0.88. Vậy tăng thêm yếu tố thì độ chính xác lên khoảng 88%.

Có thể thấy, thêm yếu tố khác vào càng nhiều thì độ dự đoán chính xác càng cao hơn, tăng lên rất rõ. Như vậy có thể dựa vào các yếu tố nhiệt độ, độ ẩm, tốc độ gió, áp suất có thể dự đoán rằng thời tiết có mưa hay là không, và từ đó có thể tìm ra dự đoán khác và có cách phòng tránh như dự đoán bão, …

Nhưng nếu như huấn luyện toàn bộ các thuộc tính thì kết quả mô hình lại ra khác:

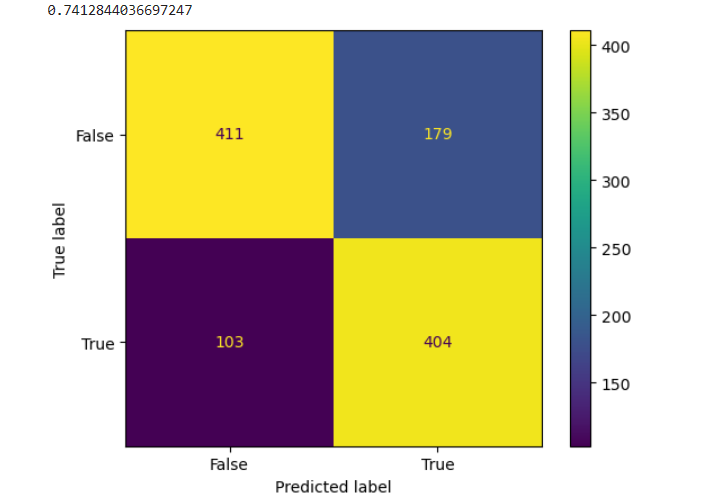
****

Mô hình 3

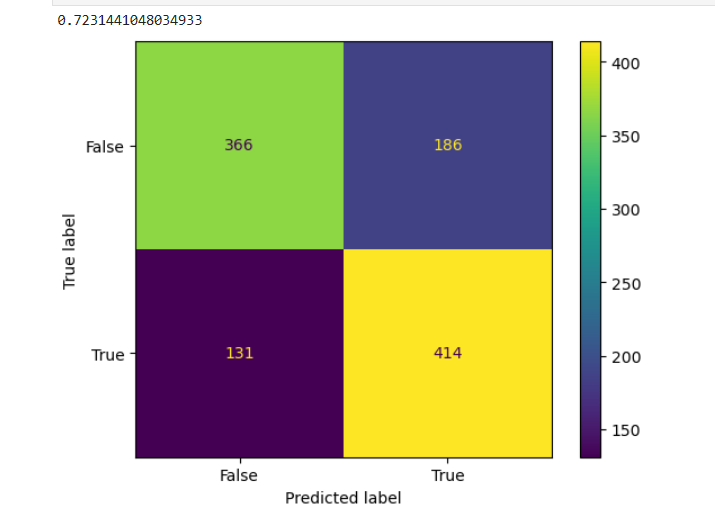
Kết quả của mô hình trên:

| True positives | True negatives | False positives | False negatives | Precision | Recall | F1 score |
| --- | --- | --- | --- | --- | --- | --- |
| 526 | 571 | 0 | 0 | 1 | 1 | 1 |

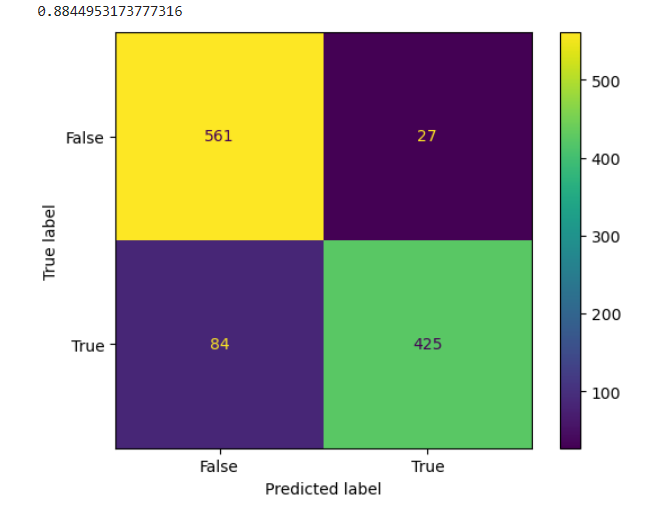
Kết quả F1-score bằng 1 cho thấy mô hình đang không đúng, bởi vì độ chính xác của mô hình là 100% cho thấy mô hình đang bị học vẹt chứ không phải dự đoán chính xác.

Sử dụng thêm Hồi quy logistic để đánh giá thêm về cây quyết định thì có các mô hình sau:

Mô hình 4

 Mô hình 5

Đây là 2 mô hình sử dụng 3 yếu tố đánh giá là độ che phủ của mây, nhiệt độ, tốc độ gió.

Mô hình 6 có tất cả các thuộc tính:

Kết quả của 3 mô hình trên:

|  | True positives | True negatives | False positives | False negatives | Precision | Recall | F1 score |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model 4 | 404 | 411 | 103 | 179 | 0.8 | 0.69 | 0.74 |
| Model 5 | 414 | 366 | 131 | 186 | 0.76 | 0.69 | 0.72 |
| Model 6 | 425 | 561 | 84 | 27 | 0.83 | 0.94 | 0.88 |

Độ chính xác của 3 mô hình:

|  | Model 4 | Model 5 | Model 6 |
| --- | --- | --- | --- |
| Độ Chính Xác | 0.74 | 0.71 | 0.89 |

Dựa vào 3 mô hình sử dụng Hồi quy logistic thì so với cùng 3 yếu tố thì mô hình có sử dụng hồi quy Logistic(mô hình 4 &5) độ chính xác thấp hơn với mô hình không sử dụng Hồi quy logistic(mô hình 1&2). Nhưng với mô hình sử dụng toàn bộ thuộc tính mô hình 6 so với mô hình 3 thì mô hình 6 có sự chính xác là 0.89 độ chính xác cao và mô hình 6 không bị tình trạng học vẹt như mô hình 3.

Vì vậy mô hình Hồi quy logistic phù hợp với dữ liệu lớn, nhiều thuộc tính hơn và đa dạng dữ liệu.

# KẾT LUẬN – HƯỚNG PHÁT TRIỂN

## Kết quả đạt được

Sau khi thực hiện đồ án môn học, chúng em đã củng cố lại những kiến thức về học máy và phân tích dữ liệu, biết phân tích dự đoán bằng thuật toán và mô hình f1-score, Hồi quy Logistic.

## Hạn chế của đồ án

Vì sự hạn hẹp về kiến thức nên nhóm chúng em vẫn còn các thiếu sót. Tuy nhiên, luận văn vẫn còn gặp một số khó khăn cần giải quyết trong thời gian tới như: Trước hết là về mặt dữ liệu. Do đây là dữ liệu kaggle, cho nên phần thu nhập dữ liệu không thể quản lý được và chưa phong phú. Dữ liệu còn hạn chế chuỗi thời gian liên tục của dữ liệu. Điều này dẫn tới sự thiếu đa dạng của dữ liệu sẽ khó lấy ra kết quả chính xác. F1-score chưa khai thác các đặc trưng khác và chưa đánh giá toàn diện các phương pháp thử nghiệm.

## Hướng phát triển

Trong thuật toán này, nhóm chúng em có thể thấy tính ưu việt của mô hình học kết hợp. Trong tương lai em sẽ tìm các bộ dữ liệu lớn hơn, có nhiều biến động hơn, để áp dụng các thuật toán một các hợp lý.

**Tài liệu tham khảo**

1 . [Weather Prediction (kaggle.com)](https://www.kaggle.com/datasets/thedevastator/weather-prediction/)

2. Slide tài liệu phân tích dữ liệu

3. Python Data Science Handbook by Jake VanderPlas

4. https://www.miai.vn/2020/06/16/oanh-gia-model-ai-theo-cach-mi-an-lien-chuong-2-precision-recall-va-f-score/?fbclid=IwAR1rIgyw29ANeSrwa7xD4shhCCpx--jC-ylqbfvNKk-AmrrTDLEPBkdQ7Wo