BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN

KHÓA LUẬN TỐT NGHIỆP

Dự báo thời tiết của TP.HCM

Giảng viên hướng dẫn: **TH.S THÁI TRÚC NHI**
Sinh viên thực hiện: **DOÃN BÙI HÒA HỢP**
MSSV: **2100005123**
Khoá: **2021 – 2025**
Ngành/ chuyên ngành: **TRÍ TUỆ NHÂN TẠO**

Tp.HCM, tháng 12 năm 2024

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN

KHÓA LUẬN TỐT NGHIỆP

Dự báo thời tiết của TP.HCM

Giảng viên hướng dẫn: **TH.S THÁI TRÚC NHI**
Sinh viên thực hiện: **DOÃN BÙI HÒA HỢP**
MSSV: **2100005123**
Khoá: **2021 - 2025**
Ngành/ chuyên ngành: **TRÍ TUỆ NHÂN TẠO**

TP.HCM, tháng 12 năm 2024

LỜI CẢM ƠN

Thưa thầy cô, khóa luận tốt nghiệp chuyên ngành Trí Tuệ Nhân Tạo với đề tài Dự Báo Thời Tiết của TP.HCM mà em vừa trình bày chính là kết quả của cả một quá trình trau dồi và nỗ lực không ngừng của bản thân em. Em rất may mắn khi luôn nhận được sự giúp đỡ, hỗ trợ tận tình từ quý thầy cô, gia đình và bạn bè của mình.

Qua đây, em xin được gửi lời cảm ơn chân thành nhất đến tất cả mọi người đã luôn giúp đỡ, động viên và khích lệ em trong suốt khoảng thời gian vừa qua.
Em xin trân trọng cảm ơn cô TH.S THÁI TRÚC NHI

Cô là người đã luôn tận tình chỉ dạy, dẫn dắt và tạo điều kiện hết sức để giúp em có thể hoàn thành tốt nhất bài luận văn của mình. Bên cạnh đó, em cũng xin được cảm ơn ban lãnh đạo cùng toàn thể các thầy cô trường Công Nghệ Thông Tin đã luôn giúp đỡ và ở bên cạnh chúng em trong suốt 3,5 năm đại học vừa qua.

Em xin một lần nữa chân thành gửi lời cảm ơn đến Khoa và các giảng viên bộ môn đã giúp em hoàn thành tốt các môn khi học tập tại trường Đại Học Nguyễn Tất Thành.

LỜI MỞ ĐẦU

Dự báo thời tiết là quá trình sử dụng các dữ liệu khí tượng và mô hình toán học để dự đoán các điều kiện thời tiết trong tương lai tại một khu vực cụ thể. Quá trình này bao gồm việc thu thập dữ liệu từ các nguồn khác nhau như vệ tinh, radar, trạm khí tượng và các cảm biến môi trường. Những dữ liệu này thường bao gồm nhiệt độ, độ ẩm, áp suất không khí, tốc độ và hướng gió, cũng như lượng mưa và các hiện tượng thời tiết khác.

Đề tài này tập chung vào việc phân tích và dự báo thời tiết cho thành phố Thành phố Hồ Chí Minh, một trong những thành phố lớn và phát triển nhất Việt Nam. Thời tiết tại TP.HCM thường xuyên biến đổi, ảnh hưởng đến đời sống hàng ngày của người dân cũng như các hoạt động kinh tế, xã hội.

Hy vọng rằng nghiên cứu này sẽ đóng góp một phần nhỏ vào sự hiểu biết của chúng ta về bản chất phức tạp của con người, cung cấp thông tin hữu ích cho các lĩnh vực như xã hội học, tâm lý học, và khoa học máy tính. Và hơn thế nữa, em sẽ cố gắng phát triển tốt lĩnh vực vào vào những nghiên cứu khác nhau để có thể tìm hiểu sâu sắc hơn đề tài nghiên cứu này.

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Hình thức

Sinh viên đã trình bày khóa luận với bố cục rõ ràng, hợp lý, tuân thủ đúng yêu cầu về mặt hình thức. Các chỉ mục, hình vẽ, bảng biểu và hình ảnh được sắp xếp khoa học, dễ theo dõi, đảm bảo tính thẩm mỹ. Phần kết luận có tóm tắt được các nội dung đã làm so với nhiệm vụ đăng ký khoá luận, có đề xuất hướng phát triển.

2. Nội dung

Sinh viên đã xác định rõ ràng mục tiêu nghiên cứu và xây dựng phương pháp thực hiện phù hợp, đảm bảo tính logic, khoa học. Nội dung các chương được trình bày chi tiết, liên kết chặt chẽ với nhau, thể hiện sự hiểu biết về vấn đề nghiên cứu. Kết quả đạt được có tính thuyết phục và ý nghĩa thực tiễn cao.

3. Kết luận

Kính đề nghị Khoa Công nghệ thông tin cho sinh viên được bảo vệ kết quả trước Hội đồng khoá luận tốt nghiệp.

TPHCM, Ngày 13 tháng 1 năm 2024

Giáo viên hướng dẫn

ThS. Thái Trúc Nhi

NHẬN XÉT CỦA GIẢNG VIÊN PHẢN BIỆN

1. Hình thức (Bố cục, trình bày, lỗi, các mục, hình, bảng, công thức, phụ lục,) .....

.....

.....

.....

.....

.....

.....

.....

.....

2. Nội dung (mục tiêu, phương pháp, kết quả, sao chép, các chương, tài liệu,).....

.....

.....

.....

.....

.....

.....

3. Kết luận.....

.....

TPHCM, Ngày 22 tháng 12 năm 2024

Giáo viên phản biện

(Ký tên, ghi rõ họ tên)

MỤC LỤC

DANH MỤC HÌNH

CHƯƠNG I: GIỚI THIỆU VỀ ĐỀ TÀI

1.1 Giới thiệu về đề tài

Trong thế giới hiện đại, thời tiết không chỉ đơn thuần là một yếu tố tự nhiên mà còn là một phần quan trọng ảnh hưởng đến mọi khía cạnh của cuộc sống con người. Từ nông nghiệp, giao thông, đến du lịch và quản lý thiên tai, việc hiểu rõ và dự đoán thời tiết chính xác trở thành một nhu cầu thiết yếu. Đề tài "Phân Tích và Dự Đoán Thời Tiết Dựa Trên Dữ Liệu Khí Tượng" nhằm mục tiêu khai thác và phân tích dữ liệu thời tiết để đưa ra những dự đoán chính xác về điều kiện khí hậu trong tương lai.

1.2 Lý do chọn đề tài

Lý do chọn đề tài dự báo thời tiết tại TP.HCM xuất phát từ tầm quan trọng của thời tiết trong đời sống hàng ngày và các hoạt động kinh tế, xã hội của thành phố. TP.HCM, với khí hậu nhiệt đới gió mùa, thường xuyên phải đối mặt với các hiện tượng thời tiết cực đoan như mưa lớn, bão và nắng nóng, ảnh hưởng trực tiếp đến sức khỏe cộng đồng, sản xuất nông nghiệp và giao thông. Do đó, việc dự đoán chính xác các điều kiện thời tiết không chỉ giúp người dân chuẩn bị tốt hơn cho các tình huống bất ngờ mà còn hỗ trợ các cơ quan chức năng trong việc lập kế hoạch ứng phó và quản lý thiên tai.

1.3 Phạm vi nghiên cứu

Phạm vi nghiên cứu dự báo thời tiết tại TP.HCM không chỉ tập trung vào việc phân tích các yếu tố khí hậu và áp dụng các mô hình dự đoán, mà còn mở rộng đến việc đánh giá độ chính xác và ứng dụng kết quả nghiên cứu vào thực tiễn. Điều này sẽ góp phần nâng cao khả năng dự đoán thời tiết, từ đó bảo vệ an toàn cho cộng đồng và phát triển bền vững cho thành phố.

1.4 Đối tượng nghiên cứu

Dự báo thời tiết tại TP.HCM bao gồm nhiều yếu tố khí hậu quan trọng, như nhiệt độ, độ ẩm, áp suất không khí, tốc độ gió và lượng mưa. Những yếu tố này không chỉ ảnh hưởng đến điều kiện thời tiết hàng ngày mà còn tác động đến các hoạt động kinh tế, xã hội và sức khỏe cộng đồng. Việc phân tích và dự đoán các yếu tố này giúp cung cấp thông tin kịp thời và chính xác cho người dân và các cơ quan chức năng. Dữ liệu khí tượng cũng là một phần quan trọng trong nghiên cứu này. Dữ liệu lịch sử từ các trạm khí tượng và vệ tinh, cùng với dữ liệu thời gian thực từ các cảm biến, sẽ được thu thập và phân tích để xây dựng các mô hình dự đoán. Việc sử dụng dữ liệu phong phú và đa dạng sẽ giúp cải thiện độ chính xác của các dự báo thời tiết.

1.5 Phạm vi nghiên cứu

Tập trung vào dự báo thời tiết tại TP.HCM, bao gồm các yếu tố khí hậu chính như nhiệt độ, độ ẩm, lượng mưa, tốc độ và hướng gió, áp suất khí quyển, và tình trạng thời tiết (nắng, mưa, mây...). Thời gian dự báo được giới hạn trong ngắn hạn (từ 1 đến 7 ngày) để đảm bảo tính khả thi và độ chính xác của mô hình. Địa điểm nghiên cứu được giới hạn trong khu vực TP.HCM, với dữ liệu lịch sử và hiện tại được thu thập từ các trạm khí tượng hoặc các nguồn cung cấp thông tin thời tiết đáng tin cậy. Xây dựng một mô hình dự báo thời tiết hiệu quả, phù hợp với đặc thù khí hậu nhiệt đới gió mùa của TP.HCM, từ đó hỗ trợ các hoạt động sản xuất, sinh hoạt và ứng phó với biến đổi khí hậu.

1.6 Bố cục đề tài

Chương I: Giới thiệu chung về đề tài

Chương II: Cơ sở lý luận về vấn đề nghiên cứu

Chương III: Mô hình lý thuyết

Chương IV: Mô hình thực nghiệm

Chương V: Kết luận và kiến nghị

CHƯƠNG II: CƠ SỞ LÝ LUẬN VỀ VẤN ĐỀ NGHIÊN CỨU

2.1 Tổng quan về dự báo thời tiết thành phố Hồ Chí Minh

TPHCM có khí hậu nhiệt đới gió mùa, với hai mùa chính: mùa mưa và mùa khô. Sự thay đổi thất thường của thời tiết, đặc biệt do biến đổi khí hậu, đặt ra thách thức lớn cho công tác dự báo. Các hiện tượng như mưa lớn gây ngập lụt, nắng nóng kéo dài ảnh hưởng đến cuộc sống và nền kinh tế. Vì vậy, việc áp dụng các phương pháp dự báo hiện đại là cần thiết để hỗ trợ quản lý thiên tai và phát triển bền vững.

2.2 Các lý thuyết và nghiên cứu liên quan

2.2.1 Lý thuyết về Dự báo thời tiết tại TP.HCM

Trong lĩnh vực khí tượng học, việc sử dụng các mô hình để phân tích và dự đoán thời tiết là một phần không thể thiếu. Các mô hình này giúp các nhà nghiên cứu và chuyên gia khí tượng hiểu rõ hơn về các yếu tố khí hậu, từ đó đưa ra các dự đoán chính xác và kịp thời. Dưới đây là một cái nhìn sâu sắc về các loại mô hình thường được sử dụng trong phân tích dữ liệu khí tượng.

2.2.2 Nghiên cứu về Dự báo thời tiết TP.HCM

+ Nghiên cứu của các tổ chức khí tượng: Nhiều tổ chức khí tượng, như Trung tâm Dự báo Khí tượng Thủy văn Quốc gia, đã thực hiện các nghiên cứu về dự đoán thời tiết tại TP.HCM, sử dụng dữ liệu từ các trạm khí tượng và vệ tinh.

Nghiên cứu về biến đổi khí hậu: Các nghiên cứu đã chỉ ra rằng biến đổi khí hậu có thể ảnh hưởng đến các yếu tố thời tiết tại TP.HCM, như nhiệt độ và lượng mưa. Việc hiểu rõ tác động của biến đổi khí hậu là rất quan trọng trong việc dự đoán thời tiết.

2.2.3 Các tập dữ liệu phổ biến của bài toán dự báo thời tiết

+ Mục Tiêu Dự Đoán:

Dự đoán các yếu tố thời tiết như nhiệt độ (Temp), độ ẩm (Humidity), lượng mưa (Rain), và điều kiện thời tiết (Weather) cho các thời điểm trong tương lai dựa trên dữ liệu lịch sử.

+ **Các Biến Độc Lập**:

Các yếu tố như thời gian (Time), ngày (Date), gió (Wind), áp suất (Pressure), và tình trạng mây (Cloud) có thể được sử dụng làm biến độc lập để dự đoán các yếu tố mục tiêu.

Phương Pháp Dự Đoán:

Có thể áp dụng các mô hình thống kê như Linear Regression hoặc các mô hình học máy như Random Forest, Support Vector Machines (SVM), và mạng nơ-ron (RNN, LSTM) để xây dựng mô hình dự đoán.

Đánh Giá Mô Hình:

Sử dụng các chỉ số như R^2, Mean Squared Error (MSE) để đánh giá độ chính xác của mô hình dự đoán.

Ứng Dụng Kết Quả:

Kết quả dự đoán có thể được sử dụng để cung cấp thông tin cho người dân và các cơ quan chức năng, giúp họ chuẩn bị tốt hơn cho các điều kiện thời tiết trong tương lai.

2.3 Vấn đề cần giải quyết và giải pháp đề xuất

2.3.1 Vấn đề cần giải quyết

Độ Chính Xác Của Dự Đoán: Độ chính xác của các mô hình dự đoán thời tiết hiện tại còn thấp, dẫn đến việc không thể cung cấp thông tin kịp thời và chính xác cho người dân và các cơ quan chức năng.

Dữ Liệu Thiếu và Không Đầy Đủ: Việc thu thập dữ liệu khí tượng có thể gặp khó khăn do thiếu các trạm quan trắc hoặc dữ liệu không đầy đủ, ảnh hưởng đến khả năng phân tích và dự đoán.

Mô Hình Dự Đoán Phù Hợp: Việc lựa chọn mô hình dự đoán phù hợp với điều kiện khí hậu tại TP.HCM là một thách thức, cần phải thử nghiệm và so sánh nhiều mô hình khác nhau.

Tích Hợp Dữ Liệu Thời Gian Thực: Cần phát triển các phương pháp để xử lý và phân tích dữ liệu thời gian thực một cách hiệu quả, giúp cải thiện khả năng dự đoán.

Ứng Phó Với Các Hiện Tượng Thời Tiết Cực Đoan: TP.HCM thường xuyên phải đối mặt với các hiện tượng thời tiết cực đoan như bão, mưa lớn và nắng nóng, cần có hệ thống cảnh báo sớm và thông tin kịp thời.

2.3.2 Giải pháp đề xuất

Cải Thiện Mô Hình Dự Đoán: Nghiên cứu và áp dụng các mô hình học máy tiên tiến như Random Forest, LSTM, và các mô hình kết hợp để nâng cao độ chính xác của dự đoán. Thực hiện kiểm tra chéo và đánh giá mô hình để chọn ra mô hình tốt nhất cho từng yếu tố thời tiết.

Tăng Cường Thu Thập Dữ Liệu: Thiết lập thêm các trạm quan trắc khí tượng và sử dụng công nghệ IoT để thu thập dữ liệu từ nhiều nguồn khác nhau. Kết hợp dữ liệu từ các trạm khí tượng, vệ tinh và cảm biến để đảm bảo tính đầy đủ và chính xác của dữ liệu.

Phát Triển Hệ Thống Cảnh Báo Sớm: Xây dựng hệ thống cảnh báo sớm cho các hiện tượng thời tiết cực đoan, sử dụng dữ liệu thời gian thực và mô hình dự đoán để cung cấp thông tin kịp thời cho người dân và các cơ quan chức năng.

Những giải pháp này sẽ giúp nâng cao độ chính xác của dự đoán thời tiết và cung cấp thông tin hữu ích cho cộng đồng và các cơ quan chức năng trong việc ứng phó với các điều kiện thời tiết.

CHƯƠNG III: MÔ HÌNH LÝ THUYẾT

3.1 Linear Regression

+ Mục Tiêu Dự Đoán

- Mô hình Linear Regression sẽ được sử dụng để dự đoán các yếu tố thời tiết như:
- Nhiệt độ (Temp)
- Độ ẩm (Humidity)
- Lượng mưa (Rain)
- Áp suất không khí (Pressure)

+ Dữ Liệu Cần Thiết

- Dữ liệu lịch sử: Sử dụng dữ liệu từ tệp weatherHCM.csv, bao gồm các yếu tố như nhiệt độ, độ ẩm, áp suất, tốc độ gió, và lượng mưa trong một khoảng thời gian dài.
- Biến độc lập: Các yếu tố như độ ẩm, áp suất không khí, tốc độ gió, và thời gian (ngày, giờ) sẽ được sử dụng làm biến độc lập để dự đoán các yếu tố mục tiêu.

+ Tiền Xử Lý Dữ Liệu

- Làm sạch dữ liệu: Xử lý các giá trị thiếu và ngoại lệ để đảm bảo tính chính xác của dữ liệu.
- Chuẩn hóa dữ liệu: Chuẩn hóa hoặc chuẩn hóa các biến để đảm bảo rằng tất cả các biến đều có cùng quy mô, giúp cải thiện hiệu suất của mô hình.
- Chia dữ liệu: Chia dữ liệu thành các tập huấn luyện và kiểm tra để đánh giá độ chính xác của mô hình.

3.2 Học máy (Linear Regression)

3.2.1 Linear Regression là gì?

Linear Regression là một thuật toán học có giám sát (supervised learning) trong Machine Learning, nó là một phương pháp thống kê dùng để ước lượng mối quan hệ giữa các biến độc lập (input features) và biến phụ thuộc (output target). Linear Regression giả định rằng sự tương quan giữa các biến là tuyến tính, từ đó tìm ra hàm tuyến tính tốt nhất để biểu diễn mối quan hệ này. Thuật toán này dự báo giá trị của biến output từ các giá trị của các biến đầu vào.Mô hình hồi quy tuyến tính có dạng:

$y = \beta 0 + \beta 1x$

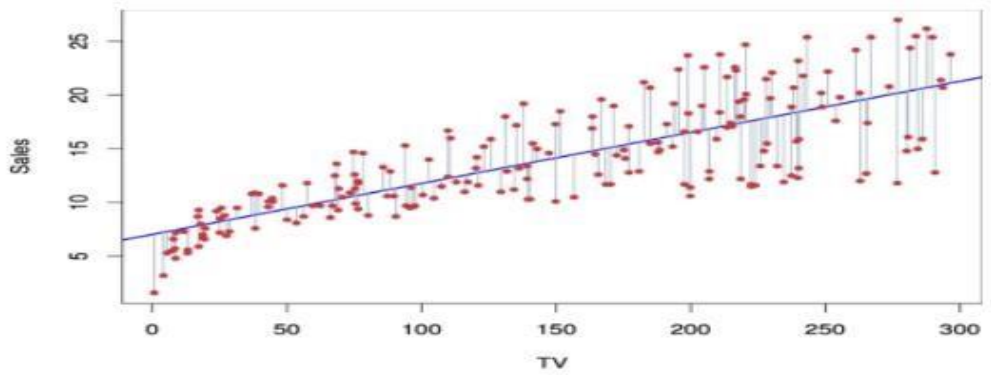
Trong đó:

- y là biến phụ thuộc
- x là biến độc lập
- β0 là hệ số chặn
- β1 là hệ số góc

Hệ số chặn β0: Hệ số chặn đại diện cho giá trị của biến phụ thuộc khi biến độc lập bằng 0.

Hệ số góc β1: Hệ số góc đại diện cho độ dốc của đường hồi quy.

Đường hồi quy tuyến tính: Đường hồi quy tuyến tính là đường đi qua các điểm dữ liệu sao cho tổng bình phương các sai số giữa giá trị dự đoán và giá trị thực tế của biến phụ thuộc là nhỏ nhất.



3.2.2 Các loại Linear Regression

Có hai loại chính của Linear Regression:

+ Simple Linear Regression: Mô hình này chỉ có một biến độc lập (input feature) mô tả mối quan hệ tuyến tính giữa biến phụ thuộc (output target) và biến độc lập. Phương trình của Simple Linear Regression có dạng y=a+bx+ϵy=a+bx+ϵ, trong đó a là điểm giao với trục tung (chỉ số độc lập), b là hệ số góc (độ dốc) của đường thẳng, và ϵϵ là sai số.

+ Multiple Linear Regression: Mô hình này có nhiều hơn một biến độc lập, biểu diễn mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc. Phương trinh của Multiple Linear Regression dạng :

y=a+b1x1+b2x2+…+bnxn+ϵy=a+b1x1+b2x2+…+bnxn+ϵ, trong đó a là điểm giao với trục tung, b1,b2,…,bnb1,b2,…,bn là các hệ số góc, và ϵϵ là sai số.

3.2.3 Mô hình hồi quy tuyến tính

Cách xây dựng mô hình hồi quy tuyến tính:

- Thu thập dữ liệu:”Dữ liệu cần bao gồm biến phụ thuộc và biến độc lập.”
- Tính toán hệ số β0 và β1: Có thể sử dụng các phương pháp như phương pháp bình phương tối thiểu, phương pháp trọng số least squares, hoặc phương pháp giải hệ phương trình tuyến tính.
- Đánh giá mô hình: Có thể sử dụng các phương pháp như kiểm định t, kiểm định F, hoặc phân tích hồi quy.

3.2.4 Mục tiêu của Linear Regression

Mục tiêu của Linear Regression là tìm ra hệ số góc và điểm giao với trục tung sao cho hàm dự đoán tuyến tính đạt được sai số nhỏ nhất. Một trong những cách phổ biến để ước lượng các hệ số là sử dụng phương pháp Ordinary Least Squares (OLS), trong đó chúng ta cần tối thiểu hóa tổng bình phương sai số (sum of squared error)

3.2.5 Ứng dụng của Linear Regression

Linear Regression được ứng dụng rộng rãi trong nhiều lĩnh vực, như:

Dự báo giá cả: dự đoán giá nhà, giá cổ phiếu, giá nhiên liệu dựa trên các yếu tố như vị trí, kích thước, chất lượng, lượng cung cầu, …

Dự báo điểm số: dự đoán điểm số của học sinh dựa trên thời gian học, nỗ lực, kỹ năng, trình độ giáo viên, …

Dự báo sản phẩm: dự đoán đầu ra sản xuất dựa trên thời gian, công suất, nguyên liệu, lao động, …

Phân tích chuỗi thời gian: dự đoán xu hướng và chu kỳ của các chuỗi dữ liệu, như bất động sản, thời tiết, xu hướng sản xuất, …

3.3 Feature Engineering

3.3.1 Feature Engineering là gì

Một quá trình quan trọng trong học máy (machine learning) và phân tích dữ liệu. Nó liên quan đến việc tạo ra, biến đổi, hoặc chọn lựa các tính năng (features) từ dữ liệu thô để cải thiện hiệu suất của mô hình học máy. Feature engineering không chỉ giúp mô hình dễ dàng học được mối quan hệ giữa dữ liệu và kết quả mà còn có thể làm cho mô hình hoạt động hiệu quả hơn với ít tài nguyên hơn.

3.3.2 Các loại tính năng Feature Engineering

Tính năng (Feature): Là các đặc điểm hoặc thuộc tính được trích xuất từ dữ liệu thô mà mô hình học máy có thể sử dụng để học và đưa ra dự đoán.

Feature Extraction (Trích xuất tính năng): Quá trình trích xuất các đặc điểm có ý nghĩa từ dữ liệu thô. Ví dụ: trích xuất các thông tin như ngày tháng từ một chuỗi thời gian.

Feature Transformation (Biến đổi tính năng): Thay đổi hoặc biến đổi các tính năng để mô hình dễ dàng học được. Ví dụ: chuẩn hóa dữ liệu, sử dụng log-transform, hoặc sử dụng PCA (Principal Component Analysis) để giảm chiều dữ liệu.

Feature Selection (Lựa chọn tính năng): Lựa chọn các tính năng quan trọng nhất cho mô hình và loại bỏ các tính năng không có giá trị hoặc có tính tương quan cao (multicollinearity) với các tính năng khác.

3.3.3 Ứng dụng Feature Engineering

+ Dự báo và phân tích dữ liệu thời gian

+ Dự báo thời tiết: Feature Engineering có thể trích xuất các đặc trưng như chu kỳ ngày/đêm, mùa trong năm, hay các chỉ số khí hậu khác để cải thiện dự đoán về nhiệt độ, độ ẩm, và các hiện tượng thời tiết khác.

3.4 Các thư viện cho bài toán dự báo thời tiết TP.HCM

3.4.1 Thư viện numpy

- Là một thư viện quan trọng trong ngôn ngữ lập trình Python, chuyên về xử lý mảng và ma trận nhiều cũng như cung cấp các hàm toán học cơ bản để làm việc với dữ liệu số. Numpy giúp tối ưu hoá và tăng hiệu suất cho các phép toán số dữ liệu mảng.

- Đối Tượng Chính: Mảng NumPy (ndarray) là đối tượng chính trong thư viện này. Đây là một cấu trúc dữ liệu mảng nhiều chiều, linh hoạt và hiệu quả, cho phép lưu trữ dữ liệu số.

1/ Vector và Ma Trận:

- Vector được xem là mảng một chiều – tập hợp của các số hoặc ký tự được sắp xếp theo một chiều Còn ma trận được xem là mảng hai chiều – tập hợp các số hoặc ký tự được xếp thành các hàng (row) và cột (columns).

$$\vec{a}(x_1, x_2, x_3), \vec{b}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

2/ Phép Toán Element-wise:

- NumPy hỗ trợ các phép toán element-wise trên mảng, nghĩa là các phép toán được áp dụng cho từng phần tử riêng lẻ của mảng mà không cần sử dụng vòng lặp.

- Phép toán Element-wise (hoặc phép toán theo phần tử) là một khái niệm quan trọng trong lĩnh vực xử lý mảng và ma trận. Trong ngữ cảnh của thư viện

Numpy, phép toán có khả năng thực hiện các phép tính trên từng phần tử riêng lẻ của mảng mà không cần sử dụng vòng lặp .

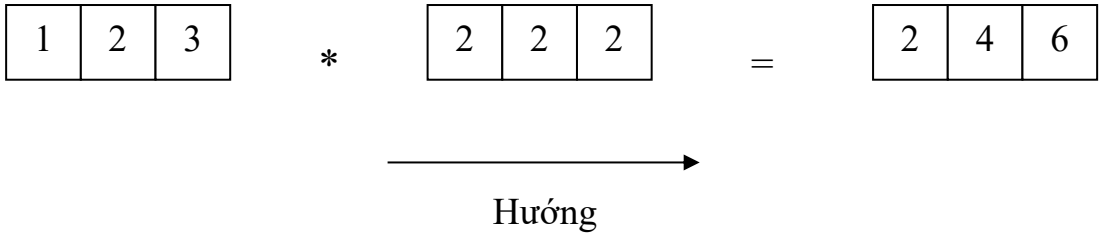
- Hiệu suất: giúp tối ưu hoá hiệu suất bằng cách áp dụng phép toán cho toàn bộ mảng một lần, thay vì thực hiện phép toán trên từng phần tử một cách tuần tự.

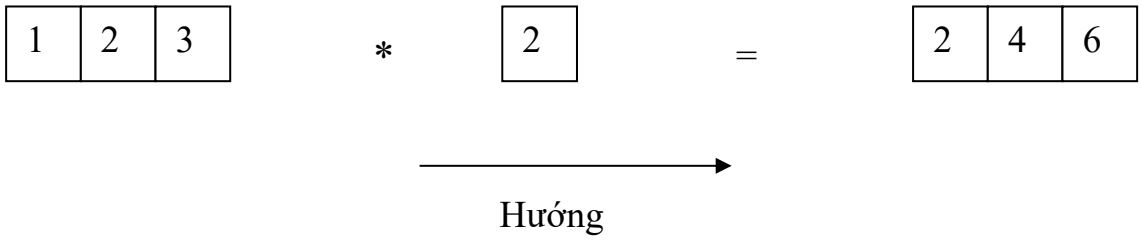
3/ Broadcasting trong Numpy

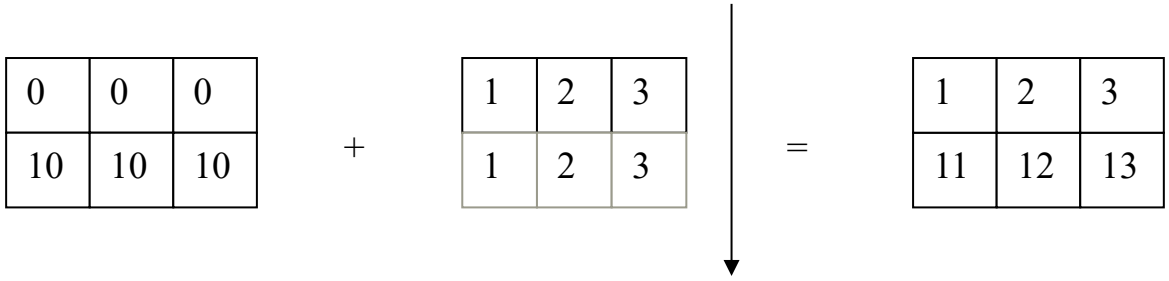
- Broadcasting: Broadcasting là khả năng của NumPy để thực hiện các phép toán giữa các mảng có kích thước khác nhau mà không cần phải tạo ra các bản sao của chúng. Broadcasting làm cho mã nguồn trở nên ngắn gọn và dễ đọc, đồng thời tăng hiệu suất bằng cách trách việc phải sao chép dữ liệu nhiều lần.

- Nguyên Tắc Cơ Bản của Broadcasting:

- Kích Thước (Shapes) Phù Hợp: Các chiều của hai mảng được coi là phù hợp nếu chúng có cùng kích thước hoặc một trong những chiều không tồn tại.
- Chiều thực hiện phép toán: Broadcasting sẽ được thực hiện theo các chiều của mảng có kích thước bằng 1. Nếu một mảng có chiều có kích thước là 1, nó sẽ được mở rộng để có kích thước giống với mảng khác.







4/ Phép Toán Hợp và Giao: là các phép toán set giữa các mảng NumPy. Các phép toán này giúp xác định các phần tử chung và khác nhau giữa hai mảng. Dưới đây là mô tả về phép toán hợp và giao trong NumPy:

- Phép toán hợp (union): Phép toán hợp giữa hai mảng trả về một mảng mới chứa tất cả các phần tử duy nhất từ cả hai mảng, mỗi phần tử chỉ xuất hiện một lần.
- Phép toán giao(intersection): Phép toán giao giữa hai mảng trả về một mảng mới chứa các phần tử chung giữa cả hai mảng.

- Lưu Ý Quan Trọng:

- Cả hai phép toán hợp và giao trả về một mảng đã được sắp xếp theo thứ tự tăng dần.
- Cả hai phép toán đều không thay đổi mảng ban đầu.
- Các phần tử trong kết quả của phép toán hợp không bao gồm các phần tử trùng lặp, mỗi phần tử chỉ xuất hiện một lần.
- Các phần tử trong kết quả của phép toán giao là các phần tử chung giữa hai mảng.

- Một Số Ứng Dụng:

- Trong xử lý dữ liệu, phép toán hợp và giao có thể được sử dụng để xác định các phần tử duy nhất hoặc chung giữa các tập dữ liệu.
- Trong thống kê, phép toán hợp và giao có thể được sử dụng để xác định các phần tử thuộc các tập con khác nhau của các tập dữ liệu.

3.4.2 Thư viện Pandas

- Thư viện pandas là một thư viện mã nguồn mở trong ngôn ngữ lập trình Python, được sử dụng rộng rãi trong xử lý và phân tích dữ liệu. Thư viện này cung cấp các cấu trúc dữ liệu linh hoạt và hiệu quả, như Series và DataFrame, để làm cho việc làm việc với dữ liệu dễ dàng và mạnh mẽ.

1/ Cấu trúc dữ liệu cơ bản trong pandas:

a. Series:

- Là một cấu trúc dữ liệu một chiều (1D), giống như một mảng hoặc danh sách.
- Có thể chứa dữ liệu của bất kỳ kiểu dữ liệu nào.
- Có chỉ số (index) gắn liền với mỗi phần tử, giúp tham chiếu và truy cập dữ liệu một cách dễ dàng.

b. DataFrame:

- Là một cấu trúc dữ liệu hai chiều (2D), giống như một bảng.
- Được tạo từ nhiều Series, mỗi Series trở thành một cột trong DataFrame.
- Có thể chứa dữ liệu của bất kỳ kiểu dữ liệu nào.

- Có cả hàng và cột được đánh số và có thể được tham chiếu bằng tên hoặc chỉ số.

2/ Cơ chế hoạt động cơ bản:

a. Tạo Đối Tượng pandas:

- Sử dụng hàm pd.Series() để tạo Series.
- Sử dụng hàm pd.DataFrame() để tạo DataFrame.

b. Thao tác và truy cập dữ liệu:

- Thực hiện các thao tác như chọn, cập nhật, xóa dữ liệu thông qua chỉ số hoặc tên cột/hàng.
- Sử dụng các phương thức như loc[] và iloc[] để thao tác với dữ liệu dựa trên chỉ số và vị trí.

c. Đọc và ghi dữ liệu:

- Sử dụng các hàm như pd.read_csv(), pd.read_excel() để đọc dữ liệu từ các nguồn khác nhau. Ví dụ như hình dưới, ta sẽ đọc bộ dữ liệu đã được huấn luyện học máy phân tích: df = pd.read_csv("data/emotion_dataset_2.csv")

Date	Time	Weather	Temp	Feels	Wind	Gust	Rain	Humidity	Cloud	Pressure	Vis
Thu 01, Jan 2009	0:00	Fog	23 °c	25 °c	9 km/h from NNW	15 km/h	0.0 mm	97%	100%	1010 mb	Poor
Thu 01, Jan 2009	3:00	Light drizzle	22 °c	25 °c	9 km/h from NNW	13 km/h	0.4 mm	97%	84%	1010 mb	Poor
Thu 01, Jan 2009	6:00	Fog	22 °c	25 °c	6 km/h from N	8 km/h	0.0 mm	98%	100%	1011 mb	Poor
Thu 01, Jan 2009	9:00	Cloudy	27 °c	31 °c	6 km/h from NNE	7 km/h	0.1 mm	83%	64%	1011 mb	Excellent
Thu 01, Jan 2009	12:00	Partly cloudy	28 °c	34 °c	3 km/h from NE	3 km/h	0.0 mm	76%	62%	1010 mb	Excellent
Thu 01, Jan 2009	15:00	Moderate or heavy ra	27 °c	32 °c	2 km/h from NNE	4 km/h	3.1 mm	83%	74%	1009 mb	Good
Thu 01, Jan 2009	18:00	Cloudy	24 °c	27 °c	7 km/h from NW	14 km/h	0.0 mm	91%	73%	1010 mb	Excellent
Thu 01, Jan 2009	21:00	Patchy rain possible	23 °c	26 °c	10 km/h from NNE	18 km/h	1.6 mm	91%	32%	1012 mb	Excellent
Fri 02, Jan 2009	0:00	Partly cloudy	22 °c	22 °c	8 km/h from N	14 km/h	0.0 mm	91%	30%	1011 mb	Excellent
Fri 02, Jan 2009	3:00	Cloudy	21 °c	22 °c	9 km/h from N	16 km/h	0.0 mm	91%	72%	1010 mb	Excellent
Fri 02, Jan 2009	6:00	Sunny	22 °c	25 °c	7 km/h from N	10 km/h	0.0 mm	91%	23%	1011 mb	Excellent
Fri 02, Jan 2009	9:00	Cloudy	26 °c	29 °c	12 km/h from NNE	14 km/h	0.0 mm	77%	68%	1012 mb	Excellent
Fri 02, Jan 2009	12:00	Cloudy	28 °c	31 °c	8 km/h from NNE	10 km/h	0.0 mm	70%	72%	1011 mb	Excellent
Fri 02, Jan 2009	15:00	Cloudy	26 °c	29 °c	9 km/h from N	13 km/h	0.1 mm	80%	73%	1010 mb	Excellent
Fri 02, Jan 2009	18:00	Partly cloudy	24 °c	26 °c	9 km/h from NNE	16 km/h	0.0 mm	82%	49%	1011 mb	Excellent
Fri 02, Jan 2009	21:00	Cloudy	23 °c	25 °c	13 km/h from N	22 km/h	0.0 mm	82%	85%	1012 mb	Excellent
Sat 03, Jan 2009	0:00	Partly cloudy	22 °c	25 °c	12 km/h from N	20 km/h	0.0 mm	82%	47%	1012 mb	Excellent
Sat 03, Jan 2009	3:00	Light drizzle	21 °c	22 °c	9 km/h from N	16 km/h	0.6 mm	88%	84%	1011 mb	Poor
Sat 03, Jan 2009	6:00	Partly cloudy	21 °c	22 °c	3 km/h from NNE	5 km/h	0.0 mm	92%	59%	1011 mb	Excellent
Sat 03, Jan 2009	9:00	Patchy light rain	23 °c	26 °c	7 km/h from NNE	9 km/h	0.8 mm	86%	61%	1013 mb	Excellent
Sat 03, Jan 2009	12:00	Partly cloudy	24 °c	27 °c	3 km/h from N	4 km/h	0.0 mm	83%	48%	1011 mb	Excellent

HÌNH 1. 1 BỘ DỮ LIỆU WEATHER

- Sử dụng các phương thức như to_csv(), to_excel() để ghi dữ liệu ra các định dạng khác nhau.

d. Thao tác và xử lý dữ liệu: Thực hiện các thao tác xử lý dữ liệu như Data(), Time(), Weather(), Feel(), Gust (),…

	Date	Time	Weather	Temp	Feels	Wind	Gust	Rain	Humidity	Cloud	Pressure	Vis
0	Thu 01, Jan 2009	00:00	Fog	23 °c	25 °c	9 km/h from NNW	15 km/h	0.0 mm	97%	100%	1010 mb	Poor
1	Thu 01, Jan 2009	03:00	Light drizzle	22 °c	25 °c	9 km/h from NNW	13 km/h	0.4 mm	97%	84%	1010 mb	Poor
2	Thu 01, Jan 2009	06:00	Fog	22 °c	25 °c	6 km/h from N	8 km/h	0.0 mm	98%	100%	1011 mb	Poor
3	Thu 01, Jan 2009	09:00	Cloudy	27 °c	31 °c	6 km/h from NNE	7 km/h	0.1 mm	83%	64%	1011 mb	Excellent
4	Thu 01, Jan 2009	12:00	Partly cloudy	28 °c	34 °c	3 km/h from NE	3 km/h	0.0 mm	76%	62%	1010 mb	Excellent
5	Thu 01, Jan 2009	15:00	Moderate or heavy rain shower	27 °c	32 °c	2 km/h from NNE	4 km/h	3.1 mm	83%	74%	1009 mb	Good
6	Thu 01, Jan 2009	18:00	Cloudy	24 °c	27 °c	7 km/h from NW	14 km/h	0.0 mm	91%	73%	1010 mb	Excellent
7	Thu 01, Jan 2009	21:00	Patchy rain possible	23 °c	26 °c	10 km/h from NNE	18 km/h	1.6 mm	91%	32%	1012 mb	Excellent
8	Fri 02, Jan 2009	00:00	Partly cloudy	22 °c	22 °c	8 km/h from N	14 km/h	0.0 mm	91%	30%	1011 mb	Excellent
9	Fri 02, Jan 2009	03:00	Cloudy	21 °c	22 °c	9 km/h from N	16 km/h	0.0 mm	91%	72%	1010 mb	Excellent

HÌNH 1. 2 THÔNG TIN DATAFRAME

- Ví dụ như đếm các giá trị trong bộ dữ liệu có bao nhiêu đếm bao nhiêu Dự báo, như hình trên ta thấy được số liệu của (34976 dòng 12 cột).

Tích hợp với NumPy: pandas tích hợp chặt chẽ với thư viện NumPy, giúp thực hiện các phép toán element-wise và broadcasting trên dữ liệu của pandas.

3.4.3 Thư viện Matplotlib

Matplotlib là một thư viện mạnh mẽ và phổ biến trong Python để tạo các biểu đồ, đồ thị 2D và 3D, hỗ trợ phân tích và trực quan hóa dữ liệu.

Có mục tiêu đơn giản hóa tối đa công việc vẽ biểu đồ để "chỉ cần vài dòng lệnh"

Hỗ trợ rất nhiều loại biểu đồ, đặc biệt là các loại được sử dụng trong nghiên cứu hoặc kinh tế như biểu đồ dòng, đường, tần suất (histograms), phổ, tương quan, errorcharts, scatterplots,…

Cấu trúc của matplotlib gồm nhiều phần, phục vụ cho các mục đích sử dụng khác nhau

Ngoài các API liên quan đến vẽ biểu đồ, matplotlib còn bao gồm một số interface: Object-Oriented API, The Scripting Interface (pyplot), The MATLAB Interface (pylab)

Các interface này giúp chúng ta thuận tiện trong việc thiết lập chỉ số trước khi thực hiện vẽ biểu đồ

Interface pylab hiện đã không còn được phát triển hầu hết các ví dụ trong slide này đều sử dụng pyplot

Sử dụng Object-Oriented API hoặc trực tiếp các API của matplotlib sẽ cho phép can thiệp sâu hơn vào việc vẽ biểu đồ (hầu hết project sẽ không có nhu cầu này)

3.4.4 Thư viện scikit learn

- Thư viện scikit-learn là một trong những thư viện phổ biến nhất trong lĩnh vực học máy (machine learning) cho Python. Nó cung cấp một loạt các công cụ và thuật toán để thực hiện các tác vụ như phân loại, hồi quy, phân cụm, và giảm chiều dữ liệu. Scikit-learn được thiết kế với giao diện đơn giản và dễ sử dụng, giúp người dùng nhanh chóng triển khai các mô hình học máy mà không cần phải hiểu sâu về các thuật toán phức tạp. Thư viện này cũng hỗ trợ nhiều phương pháp tiền xử lý dữ liệu, như chuẩn hóa và mã hóa, giúp cải thiện hiệu suất của các mô hình. Với một cộng đồng lớn và tài liệu phong phú, scikit-learn là lựa chọn hàng đầu cho cả những người mới bắt đầu và các chuyên gia trong lĩnh vực học máy.

1/ Mô Hình Học Máy:

- Scikit-learn cung cấp các mô hình học máy cho nhiều nhiệm vụ, bao gồm:

- Học giám sát (Supervised Learning): Phân loại (Classification), Hồi quy (Regression).
- Học không giám sát (Unsupervised Learning): Phân cụm (Clustering), Phân tích thành phần chính (PCA), Giảm chiều dữ liệu (Dimensionality Reduction).

- Học bán giám sát và tăng cường (Semi-supervised and Reinforcement Learning).

2/ Giao Diện Dễ Sử Dụng:

- Scikit-learn được thiết kế với một giao diện đồng nhất và dễ sử dụng cho các mô hình khác nhau. Giao diện này giúp người dùng dễ dàng chuyển đổi giữa các mô hình và thực hiện các thử nghiệm một cách thuận tiện.

3/ Tổ Chức Dữ Liệu Đồng Nhất:

- Dữ liệu đầu vào trong scikit-learn được tổ chức thành các mảng NumPy hoặc ma trận thưa (sparse matrices), làm cho việc tích hợp với các thư viện khác như NumPy, pandas trở nên thuận lợi.

4/ Tiền Xử Lý Dữ Liệu:

- Tiền xử lý dữ liệu (preprocessing) là quá trình tiền xử lý và chuẩn bị dữ liệu trước khi đưa vào mô hình học máy. Đây là bước quan trọng để đảm bảo rằng dữ liệu đầu vào là chất lượng, phù hợp với mô hình và có thể mang lại hiệu suất tốt

a. Chuẩn Hóa Dữ Liệu (Normalization/Scaling): Chuẩn hóa dữ liệu là quá trình biến đổi các giá trị của các biến thành một phạm vi nhất định. Mục tiêu là để mô hình không bị ảnh hưởng bởi sự chênh lệch về đơn vị của các biến.

- Min-Max Scaling: Chuyển đổi giá trị của biến để nằm trong khoảng giá trị đã chọn (thường là [0, 1]).
- Standardization (Z-score Scaling): Chuyển đổi giá trị của biến để có giá trị trung bình là 0 và độ lệch chuẩn là 1.

b. Xử Lý Dữ Liệu Thiếu (Handling Missing Data):

- Dữ liệu thiếu có thể làm giảm hiệu suất của mô hình. Có một số kỹ thuật để xử lý dữ liệu thiếu.
- Xóa Dữ Liệu Thiếu (Dropping): Xóa các hàng hoặc cột chứa dữ liệu thiếu.
- Điền Dữ Liệu Thiếu (Imputation): Điền giá trị cho dữ liệu thiếu.

c. Mã Hóa Biến Phân Loại (Encoding Categorical Variables):

- Mô hình học máy thường yêu cầu dữ liệu đầu vào là số hóa. Do đó, biến phân loại (categorical variables) cần được chuyển đổi thành dạng số.
- One-Hot Encoding: Tạo các biến giả (dummy variables) cho mỗi giá trị duy nhất trong biến phân loại.

d. Loại Bỏ Biến Dư Thừa (Removing Redundant Variables):

- Nếu có các biến tuyến tính phụ thuộc lẫn nhau (multicollinearity), mô hình có thể bị ảnh hưởng. Loại bỏ các biến dư thừa có thể cải thiện hiệu suất.
- Phân tích VIF (Variance Inflation Factor): Đánh giá mức độ tuyến tính phụ thuộc giữa các biến.

e. Rút Trích Đặc Trưng (Feature Extraction):

- Rút trích đặc trưng giúp giảm chiều dữ liệu và giữ lại thông tin quan trọng.
- Principal Component Analysis (PCA): Giảm chiều dữ liệu bằng cách chuyển đổi các biến ban đầu thành các thành phần chính.

f. Chia Dữ Liệu (Data Splitting):

- Chia dữ liệu thành tập huấn luyện và tập kiểm tra để đánh giá hiệu suất của mô hình trên dữ liệu mới.

- Tiền xử lý dữ liệu là bước quan trọng trong quy trình học máy. Các kỹ thuật tiền xử lý giúp cải thiện chất lượng dữ liệu, loại bỏ nhiễu, chuẩn hóa và biến đổi dữ

5/ Đánh Giá và Tối Ưu Hóa Mô Hình:

- Đây là hai khía cạnh quan trọng trong quá trình phát triển mô hình học máy. Việc này giúp cho chúng ta lựa chọn tốt mô hình huấn luyện để phù hợp và cụ thể với dự án.

a. Hồi Quy (Regression):

- Mean Squared Error (MSE): Đo lường độ lớn của sai số bình phương trung bình giữa giá trị dự đoán và giá trị thực.
- Mean Absolute Error (MAE): Đo lường sai số tuyệt đối trung bình giữa giá trị dự đoán và giá trị thực.

b. Phân Loại (Classification):

- Accuracy: Tỉ lệ dự đoán đúng trên tổng số mẫu.
- Precision, Recall, F1-Score: Đo lường hiệu suất của mô hình trên mỗi lớp.
- Area Under the Receiver Operating Characteristic Curve (AUC-ROC): Đo lường khả năng phân biệt của mô hình phân loại nhị phân.

c. Cross-Validation:

- Sử dụng kỹ thuật chia dữ liệu thành các tập huấn luyện và tập kiểm tra để đánh giá hiệu suất của mô hình.
- Phổ biến nhất là k-fold cross-validation, trong đó dữ liệu được chia thành k phần, mô hình được huấn luyện trên k-1 phần và đánh giá trên phần còn lại. Quá trình này được lặp lại k lần, mỗi lần chọn một phần khác nhau làm tập kiểm tra.

d. Confusion Matrix: Ma trận nhầm lẫn thể hiện số lượng dự đoán đúng và sai cho từng lớp trong mô hình phân loại.

e. Learning Curves: Đồ thị thể hiện biểu đồ độ lỗi hoặc độ chính xác trên tập huấn luyện và tập kiểm tra theo thời gian hoặc số lượng mẫu.

f. Bias-Variance Tradeoff: Mối quan hệ giữa sai số do bias (độ chệch) và variance (phương sai). Mô hình có thể quá đơn giản (underfitting) hoặc quá phức tạp (overfitting).

6/ Feature Engineering:

- Feature Engineering là quá trình tạo ra hoặc chọn lọc các đặc trưng từ dữ liệu đầu vào để cải thiện hiệu suất của mô hình học máy. Nó đóng vai trò quan trọng trong quá trình phát triển mô hình, giúp mô hình hiểu biểu diễn của dữ liệu một cách tốt hơn và đồng thời giảm độ phức tạp của mô hình.

- Loại Bỏ Đặc Trưng Dư Thừa: Giảm chiều của dữ liệu - Giảm độ phức tạp của mô hình.
- Xử Lý Dữ Liệu Thiếu: Đối với mô hình học máy, cần phải xử lý dữ liệu thiếu một cách thích hợp để tránh làm suy giảm hiệu suất của mô hình.

- Encoding Biến Phân Loại: Chuyển đổi các biến phân loại thành dạng số để mô hình có thể hiểu được.
- Tạo Đặc Trưng Từ Ngày Tháng: Tạo ra các đặc trưng mới từ thông tin về ngày tháng có thể giúp mô hình hiểu được các xu hướng theo thời gian.
- Tính Toán Đặc Trưng Tổng Hợp: Tính toán các đặc trưng mới dựa trên sự kết hợp của các đặc trưng hiện có.
- Rút Trích Đặc Trưng Từ Văn Bản: Rút trích thông tin quan trọng từ văn bản để sử dụng làm đặc trưng cho mô hình.
- Điều Chỉnh Scale Các Đặc Trưng: Đảm bảo rằng các đặc trưng có cùng scale để tránh ảnh hưởng của đặc trưng lớn hơn lên quá trình huấn luyện mô hình.
- Tạo Đặc Trưng Tương Quan: Tạo ra các đặc trưng mới phản ánh mối quan hệ tương quan giữa các đặc trưng hiện có.
- Tạo Đặc Trưng Phức Tạp: Tạo ra các đặc trưng phức tạp dựa trên sự kết hợp của nhiều đặc trưng để mô phỏng các mối quan hệ phức tạp.

- Feature Engineering đòi hỏi sự hiểu biết sâu rộng về dữ liệu và vấn đề cụ thể. Kỹ thuật này có thể cải thiện độ chính xác và khả năng tổng quát hóa của mô hình, đồng thời giúp mô hình dễ hiểu và giảm overfitting.

3.4.5 Thư viện Spacy

- Spacy là một thư viện xử lý ngôn ngữ tự nhiên (NLP) trong Python, được thiết kế để thực hiện các nhiệm vụ như phân tích ngôn ngữ, tách từ, nhận diện thực thể và nhiều tác vụ NLP khác.

- Spacy được phát triển để cung cấp một cách tiếp cận hiệu quả và nhanh chóng cho các tác vụ xử lý ngôn ngữ tự nhiên.

1/ Các kỹ thuật trong Spacy

- Tokenization (Tách Từ): Spacy có một bộ tách từ thật sự hiệu quả, có khả năng xử lý nhiều loại ngôn ngữ khác nhau. Kỹ thuật này không chỉ

đơn giản là cắt văn bản thành các từ, mà còn xử lý các trường hợp như từ viết tắt, từ liên kết và các trường hợp phức tạp như tokenization.

- POS Tagging (Gán Nhãn Loại Từ Loại): Xác định loại từ (danh từ, động từ, tính từ, ...) cho mỗi token trong văn bản.
- Dependency Parsing (Phân Tích Phụ Thuộc): Xây dựng cây phụ thuộc để mô tả cấu trúc ngữ pháp của câu.
- Named Entity Recognition (Nhận Diện Thực Thể): Xác định và phân loại các thực thể như tên riêng, địa điểm, ngày tháng,…
- Word Vectors (Vector Từ): Spacy sử dụng các vector từ đặc trưng để biểu diễn từ vựng dưới dạng không gian nhiều chiều, giúp mô hình hiểu ngữ cảnh và tương quan giữa các từ.

2/ Nguyên Lý Hoạt Động:

- Nguyên lý hoạt động của Spacy dựa trên việc sử dụng các mô hình học máy tiên tiến để thực hiện các tác vụ NLP. Một số đặc điểm chính bao gồm:

- Mô Hình Học Máy Tiên Tiến: Spacy sử dụng các mô hình học máy đã được đào tạo trước (pre-trained) để thực hiện các tác vụ NLP. Các mô hình này được đào tạo trên các tập dữ liệu lớn để hiểu biểu diễn của ngôn ngữ tự nhiên.
- Pipeline Xử Lý: Spacy có một pipeline xử lý mặc định, mà bạn có thể tùy chỉnh để chọn chức năng cụ thể mà bạn muốn sử dụng. Các chức năng này có thể bao gồm tách từ, gán nhãn từ loại, phân tích phụ thuộc, và nhận diện thực thể.

3/ Công Dụng:

- Spacy được sử dụng rộng rãi trong nhiều ứng dụng NLP, bao gồm:

- Xử lý ngôn Ngữ Tự Nhiên: Phân tích, hiểu và xử lý văn bản tự nhiên.
- Trích xuất thông tin: Nhận diện và trích xuất thông tin quan trọng từ văn bản, như tên riêng, ngày tháng, địa điểm.
- Dịch ngôn ngữ: Dịch văn bản giữa các ngôn ngữ khác nhau.

- Phân tích tâm lý người dùng: Hiểu ngữ cảm và tâm lý từ văn bản.
- Chatbots và ứng dụng AI: Tích hợp vào các ứng dụng AI và chatbots để hiểu và đáp ứng tự nhiên với người dùng.
- Spacy là một công cụ mạnh mẽ cho các nhà phát triển và nhà nghiên cứu trong lĩnh vực NLP và xử lý ngôn ngữ tự nhiên.

3.5 Phân tích yêu cầu

3.5.1 Phân tích yêu cầu của bài toán

+ Xác định Mục Tiêu

Mục tiêu chính: Xác định rõ mục tiêu của bài toán. Ví dụ, bạn có thể muốn phân tích dữ liệu thời tiết để dự đoán nhiệt độ, lượng mưa, hoặc các yếu tố khí hậu khác.

Kết quả mong đợi: Đặt ra các kết quả cụ thể mà bạn muốn đạt được, chẳng hạn như một mô hình dự đoán chính xác, báo cáo thống kê, hoặc biểu đồ trực quan hóa dữ liệu.

+ Phân Tích dữ liệu

Tình trạng hiện tại: Xem xét tình trạng hiện tại của dữ liệu trong file CSV. Dữ liệu có đầy đủ và chính xác không? Có cần làm sạch dữ liệu không?

Vấn đề hiện tại: Xác định các vấn đề hoặc hạn chế trong dữ liệu mà bạn cần giải quyết, chẳng hạn như dữ liệu thiếu, dữ liệu không nhất quán, hoặc định dạng không đúng.

+ Yêu cầu dữ liêu

Liệt kê các chức năng mà hệ thống hoặc giải pháp cần có để xử lý bộ dữ liệu. Ví dụ:

- Đọc dữ liệu từ file CSV.
- Làm sạch và tiền xử lý dữ liệu.
- Phân tích dữ liệu (tính toán thống kê, tìm kiếm mẫu).
- Xuất kết quả ra file mới hoặc báo cáo.

Tính năng bổ sung: Xem xét các tính năng bổ sung có thể cải thiện trải

nghiệm người dùng, chẳng hạn như khả năng trực quan hóa dữ liệu.

3.5.2 Về bộ dữ liệu cho bài toàn

+ Đặc Điểm Bộ Dữ Liệu

Nguồn dữ liệu: Xác định nguồn gốc của bộ dữ liệu. Dữ liệu có thể được thu thập từ các cảm biến thời tiết, trang web, hoặc cơ sở dữ liệu công khai.

Định dạng dữ liệu: Bộ dữ liệu được lưu trữ dưới định dạng CSV, với các cột đại diện cho các đặc trưng khác nhau như nhiệt độ, độ ẩm, lượng mưa, v.v.

+ Các Trường Dữ Liệu

Các cột trong file CSV: Mô tả các cột trong bộ dữ liệu, chẳng hạn như:

- Date: Ngày tháng ghi nhận dữ liệu.
- Temperature: Nhiệt độ (đơn vị độ C).
- Humidity: Độ ẩm (%).
- Rain: Lượng mưa (mm).
- Wind Speed: Tốc độ gió (km/h).
- Weather: Tình trạng thời tiết (mưa, nắng, mây, v.v.).

+ Chất lượng dữ liệu

- Kiểm tra chất lượng dữ liệu: Đánh giá chất lượng của bộ dữ liệu, bao gồm việc kiểm tra dữ liệu thiếu, dữ liệu không hợp lệ, và các giá trị ngoại lệ.
- Yêu cầu làm sạch dữ liệu: Xác định các bước cần thiết để làm sạch dữ liệu, chẳng hạn như loại bỏ các hàng có dữ liệu thiếu hoặc thay thế các giá trị không hợp lệ.

CHƯƠNG IV: MÔ HÌNH THỰC NGHIỆM

4.1 Tổng quan quy trình phân tích

Mô hình lý thuyết cho mã nguồn trong toàn bộ file này, tập trung vào việc phân tích và dự đoán thời tiết tại TP.HCM dựa trên dữ liệu thời tiết lịch sử. Mô hình lý thuyết sẽ bao gồm các bước từ thu thập dữ liệu, tiền xử lý, phân tích, xây dựng mô hình dự đoán, và cuối cùng là xuất kết quả. Mục tiêu là cung cấp một cái nhìn tổng quan về cách thức hoạt động của mã và các phương pháp được sử dụng để đạt được mục tiêu phân tích.

4.2 Dữ liệu phân tích

4.2.1 Bộ dữ liệu

Date	Time	Weather	Temp	Feels	Wind	Gust	Rain	Humidity	Cloud	Pressure	Vis
Thu 01, Jan 2009	0:00	Fog	23 °c	25 °c	9 km/h from NNW	15 km/h	0.0 mm	97%	100%	1010 mb	Poor
Thu 01, Jan 2009	3:00	Light drizzle	22 °c	25 °c	9 km/h from NNW	13 km/h	0.4 mm	97%	84%	1010 mb	Poor
Thu 01, Jan 2009	6:00	Fog	22 °c	25 °c	6 km/h from N	8 km/h	0.0 mm	98%	100%	1011 mb	Poor
Thu 01, Jan 2009	9:00	Cloudy	27 °c	31 °c	6 km/h from NNE	7 km/h	0.1 mm	83%	64%	1011 mb	Excellent
Thu 01, Jan 2009	12:00	Partly cloudy	28 °c	34 °c	3 km/h from NE	3 km/h	0.0 mm	76%	62%	1010 mb	Excellent
Thu 01, Jan 2009	15:00	Moderate or heavy ra	27 °c	32 °c	2 km/h from NNE	4 km/h	3.1 mm	83%	74%	1009 mb	Good
Thu 01, Jan 2009	18:00	Cloudy	24 °c	27 °c	7 km/h from NW	14 km/h	0.0 mm	91%	73%	1010 mb	Excellent
Thu 01, Jan 2009	21:00	Patchy rain possible	23 °c	26 °c	10 km/h from NNE	18 km/h	1.6 mm	91%	32%	1012 mb	Excellent
Fri 02, Jan 2009	0:00	Partly cloudy	22 °c	22 °c	8 km/h from N	14 km/h	0.0 mm	91%	30%	1011 mb	Excellent
Fri 02, Jan 2009	3:00	Cloudy	21 °c	22 °c	9 km/h from N	16 km/h	0.0 mm	91%	72%	1010 mb	Excellent
Fri 02, Jan 2009	6:00	Sunny	22 °c	25 °c	7 km/h from N	10 km/h	0.0 mm	91%	23%	1011 mb	Excellent
Fri 02, Jan 2009	9:00	Cloudy	26 °c	29 °c	12 km/h from NNE	14 km/h	0.0 mm	77%	68%	1012 mb	Excellent
Fri 02, Jan 2009	12:00	Cloudy	28 °c	31 °c	8 km/h from NNE	10 km/h	0.0 mm	70%	72%	1011 mb	Excellent
Fri 02, Jan 2009	15:00	Cloudy	26 °c	29 °c	9 km/h from N	13 km/h	0.1 mm	80%	73%	1010 mb	Excellent
Fri 02, Jan 2009	18:00	Partly cloudy	24 °c	26 °c	9 km/h from NNE	16 km/h	0.0 mm	82%	49%	1011 mb	Excellent
Fri 02, Jan 2009	21:00	Cloudy	23 °c	25 °c	13 km/h from N	22 km/h	0.0 mm	82%	85%	1012 mb	Excellent
Sat 03, Jan 2009	0:00	Partly cloudy	22 °c	25 °c	12 km/h from N	20 km/h	0.0 mm	82%	47%	1012 mb	Excellent
Sat 03, Jan 2009	3:00	Light drizzle	21 °c	22 °c	9 km/h from N	16 km/h	0.6 mm	88%	84%	1011 mb	Poor
Sat 03, Jan 2009	6:00	Partly cloudy	21 °c	22 °c	3 km/h from NNE	5 km/h	0.0 mm	92%	59%	1011 mb	Excellent
Sat 03, Jan 2009	9:00	Patchy light rain	23 °c	26 °c	7 km/h from NNE	9 km/h	0.8 mm	86%	61%	1013 mb	Excellent
Sat 03, Jan 2009	12:00	Partly cloudy	24 °c	27 °c	3 km/h from N	4 km/h	0.0 mm	83%	48%	1011 mb	Excellent

HÌNH 1. 3 BỘ DỮ LIỆU

- Bộ dữ liệu em lấy từ ACCUWEATHER và lấy dữ liệu em làm file CSV HCM và em dựa wordweather để so sánh . Chúng ta thấy bảng dữ liệu thời tiết trên cung cấp thông tin chi tiết về điều kiện thời tiết tại một địa điểm cụ thể vào các thời điểm khác nhau trong từ ngày 01/01/2009 đến ngày 20/12/2020 và được cập nhật sau mỗi 3 giờ.

- Date: ngày
- Time: thời gian trong ngày
- Weather: tình trạng thời tiết
- Temp: nhiệt độ
- Feels: nhiệt độ thực sự cảm nhận được
- Wind: tốc độ gió

- Gust: tốc độ gió giật
- Rain: lượng mưa
- Humidity: độ ẩm
- Cloud: mật độ mây
- Pressure: áp suất không khí
- Vis: tầm nhìn

Chúng ta đặt câu hỏi Bộ dữ liệu thời tiết ghi nhận vào từ ngày 01/01/2009 đến ngày 20/12/2020 và được cập nhật sau mỗi 3 giờ, cho thấy nhiều thông tin. Thời tiết phổ biến nhất trong ngày là sương mù và mưa nhẹ, với nhiệt độ dao động từ 22°C đến 27°C. Có thể thấy rằng nhiệt độ cảm nhận thường cao hơn nhiệt độ thực tế, điều này có thể liên quan đến độ ẩm không khí cao, thường trên 80%. Tốc độ gió không có sự ảnh hưởng rõ rệt đến lượng mưa, nhưng có thể thấy rằng những thời điểm có gió mạnh hơn thường đi kèm với tình trạng thời tiết xấu hơn. Tầm nhìn tốt nhất được ghi nhận vào lúc 09:00, khi trời nhiều mây nhưng không có mưa. Cuối cùng, áp suất không khí có sự thay đổi nhẹ trong suốt cả ngày, từ 1010 mb đến 1011 mb, cho thấy sự ổn định trong điều kiện thời tiết..

4.2.2 Quá trình dữ liệu

Khi em dùng data.shape trong Python được sử dụng để kiểm tra kích thước của DataFrame data.

```
(34976, 12)
```

HÌNH 1. 4 KIỂM TRA DỮ LIỆU

Chúng ta thấy bảng dữ liệu thời tiết mà em có kích thước 34,976 dòng và 12 cột, cho thấy một lượng thông tin phong phú và đa dạng. Mỗi hàng đại diện cho một bản ghi thời tiết cụ thể, trong khi các cột chứa các thông tin chi tiết như ngày, giờ, tình trạng thời tiết, nhiệt độ, độ ẩm, và nhiều yếu tố khác. Với số lượng bản ghi lớn như vậy, bộ dữ liệu này có thể cung cấp cái

nhìn sâu sắc về xu hướng thời tiết theo thời gian ,giúp phân tích và dự đoán điều kiện trong tương lai Việc khai thác dữ liệu này có thể hỗ trợ trong nhiều lĩnh vực, từ nông nghiệp đến quy hoạch đô thị, và thậm chí là nghiên cứu về biến đổi khí hậu.

Sau đó em dựa kiểm tra thông tin cung cấp và kiểm tra kết quả các dòng dữ liệu giá trị thiếu hay giá trị không thiếu của tập csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34976 entries, 0 to 34975
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            34976 non-null  object
 1   Time            34976 non-null  object
 2   Weather         34976 non-null  int64
 3   Temp            34976 non-null  float64
 4   Feels           34976 non-null  float64
 5   Gust            34976 non-null  float64
 6   Rain            34976 non-null  float64
 7   Humidity        34976 non-null  float64
 8   Cloud           34976 non-null  float64
 9   Pressure        34976 non-null  float64
 10  Vis             34976 non-null  object
 11  Wind Speed      34976 non-null  float64
 12  Wind Direction  34976 non-null  int64
 13  Month           34976 non-null  object
dtypes: float64(8), int64(2), object(4)
memory usage: 3.7+ MB
```

HÌNH 1. 5 THÔNG TIN DỮ LIỆU

Khi xuất file ra thì không có giá trị thiếu và dữ liệu của hình thì thành công và không có giá trị thiếu . Nhưng em cần thêm những giá trị phù hợp hay không thì hàm “Date ” , “Time”có thể chuyển sang kiểu datatime và muốn xử lý thời gian như nào một cách hợp lý . Hàm dữ liệu “Month” hiện đang là object của hình và có thể thay đổi và chuyển sang dạng số hay dữ liệu hay không .

Thay vì đó có thể ở phân đoạn “Tiền xử lý dữ liệu”, nhưng ta cũng thấy rõ được là ở cột này đã được lược bỏ một số từ đáng kể và tất nhiên công dụng của

nó là để dễ dàng hơn khi xử lý những thời tiết của các đoạn cột đó và đưa ra kết quả có khi không được chính xác hơn.

4.3 Dữ liệu thống kê

4.3.1 Phân bố dữ liệu

Trong quá trình phân tích dữ liệu thời tiết, việc hiểu rõ các đặc điểm thống kê của từng biến là rất quan trọng. Bảng thống kê mô tả dưới đây cung cấp thông tin chi tiết về các biến số liên quan đến thời tiết, bao gồm nhiệt độ, độ ẩm, lượng mưa, và các yếu tố khác. Thông qua các chỉ số thống kê như trung bình, độ lệch chuẩn, và các giá trị tối thiểu/tối đa, chúng ta có thể có cái nhìn tổng quan về dữ liệu và phát hiện các xu hướng hoặc bất thường.

Sau khi chọn những mục cột trong bảng weather thì các hàng thuộc tính như Temp, Feels, Gust, Rain, và Presure. Sau đó được kết quả danh mục bảng thông số Rain, và Pressure vòng lặp for để duyệt qua :

	Temp	Feels	Gust	Rain	Humidity	Cloud	Pressure	Wind Speed
count	34976.000000	34976.000000	34976.000000	34976.000000	34976.000000	34976.000000	34976.000000	34976.000000
mean	27.845437	32.188758	13.076538	0.441763	0.739599	0.379355	1009.337917	9.156593
std	3.413878	4.438785	6.687806	1.458789	0.156772	0.243661	2.388064	4.507744
min	16.000000	16.000000	0.000000	0.000000	0.210000	0.000000	1000.000000	0.000000
25%	25.000000	29.000000	8.000000	0.000000	0.640000	0.190000	1008.000000	6.000000
50%	27.000000	31.000000	12.000000	0.000000	0.770000	0.330000	1009.000000	9.000000
75%	30.000000	36.000000	17.000000	0.000000	0.870000	0.530000	1011.000000	12.000000
max	39.000000	51.000000	55.000000	50.800000	0.990000	1.000000	1019.000000	33.000000

HÌNH 1. 6 DỮ LIỆU DẠNG SỐ

Chúng ta thấy một tập dữ liệu thời tiết với tổng cộng 34,976 điểm dữ liệu. Các thông số được phân tích bao gồm nhiệt độ (Temp), cảm giác nhiệt (Feels), tốc độ gió giật (Gust), lượng mưa (Rain), độ ẩm (Humidity), mức độ che phủ mây (Cloud), áp suất khí quyển (Pressure), và tốc độ gió (Wind Speed). Dưới đây là mô tả chi tiết hơn về từng thông số:

- Nhiệt độ (Temp): Nhiệt độ trung bình được ghi nhận là 27.85°C, với mức thấp nhất là 16°C và cao nhất là 39°C. Phân bố nhiệt độ khá hẹp với độ

lệch chuẩn 3.41°C, cho thấy phần lớn các giá trị không cách quá xa trung bình.

- Cảm giác nhiệt (Feels): Đây là giá trị thể hiện cảm nhận thực tế của con người về nhiệt độ. Trung bình, cảm giác nhiệt là 32.19°C, cao hơn nhiệt độ trung bình thực tế, có thể do ảnh hưởng của các yếu tố như độ ẩm và gió. Giá trị tối đa lên đến 51°C, cho thấy sự khắc nghiệt có thể xảy ra trong một số trường hợp.
- Gió giật (Gust): Tốc độ gió giật trung bình là 13.08 km/h, với mức cao nhất đạt tới 55 km/h. Điều này thể hiện sự biến thiên khá lớn, đặc biệt trong những trường hợp có hiện tượng thời tiết bất thường như giông bão.
- Lượng mưa (Rain): Lượng mưa trung bình được ghi nhận khá thấp, chỉ 0.44 mm. Điều này cho thấy phần lớn thời gian trong tập dữ liệu không có mưa, hoặc lượng mưa nhỏ, nhưng giá trị tối đa đạt 50.8 mm cho thấy có những thời điểm xảy ra mưa lớn.
- Độ ẩm (Humidity): Độ ẩm trung bình là 73.96%, dao động từ 21% đến 99%. Độ lệch chuẩn thấp (15.68%) cho thấy độ ẩm thường xuyên duy trì ở mức cao, phù hợp với môi trường khí hậu nhiệt đới hoặc ẩm ướt.
- Mức độ che phủ mây (Cloud): Trung bình, mức độ mây che phủ bầu trời là 37.94%. Tuy nhiên, giá trị có thể dao động từ bầu trời hoàn toàn trong xanh (0%) đến bầu trời hoàn toàn bị che phủ (100%).
- Áp suất khí quyển (Pressure): Áp suất trung bình là 1009.34 hPa, với phạm vi dao động từ 1000 hPa đến 1019 hPa. Giá trị này cho thấy điều kiện khí quyển khá ổn định, phù hợp với điều kiện thời tiết thông thường.
- Tốc độ gió (Wind Speed): Tốc độ gió trung bình là 9.16 km/h, với giá trị tối đa là 33 km/h. Mặc dù giá trị trung bình khá thấp, nhưng một số thời điểm gió có thể mạnh hơn, đặc biệt trong các tình huống thời tiết cực đoan.

Nhìn chung, bảng thống kê này cung cấp một bức tranh toàn cảnh về các yếu tố thời tiết trong tập dữ liệu. Những giá trị trung bình và độ lệch chuẩn giúp đánh giá mức độ ổn định của từng thông số, trong khi các giá trị cực trị (min và

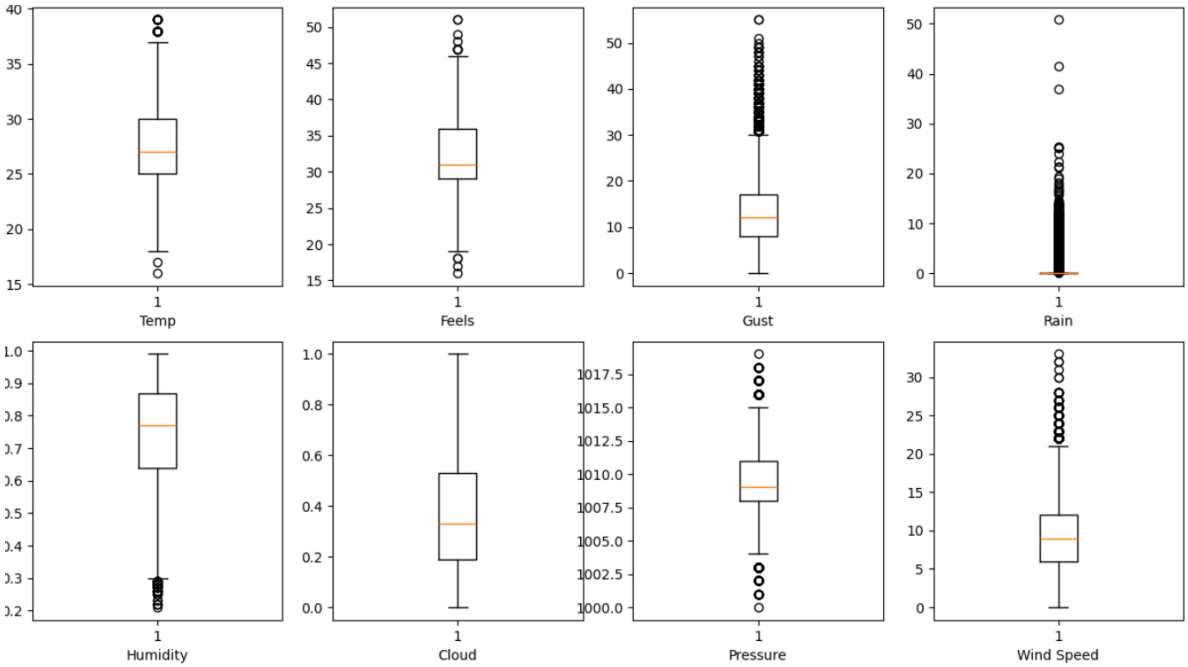
max) phản ánh các tình huống đặc biệt hoặc bất thường. Tập dữ liệu này có thể hữu ích trong việc phân tích, xây dựng mô hình dự báo thời tiết, hoặc tìm hiểu mối quan hệ giữa các yếu tố môi trường khác nhau.

Sau đó em gọi hàm fig, axs = plt.subplots nó như kiểu dạng boxplot là dạng biểu đồ thống kê sử dụng biểu diễn so sánh và quan sát dữ liệu trong nhiều biến

```
fig, axs = plt.subplots(2, 4, figsize=(15, 8))

axs[0, 0].boxplot(data['Temp'])
axs[0, 0].set_xlabel('Temp')

axs[0, 1].boxplot(data['Feels'])
axs[0, 1].set_xlabel('Feels')

axs[0, 2].boxplot(data['Gust'])
axs[0, 2].set_xlabel('Gust')

axs[0, 3].boxplot(data['Rain'])
axs[0, 3].set_xlabel('Rain')

axs[1, 0].boxplot(data['Humidity'])
axs[1, 0].set_xlabel('Humidity')

axs[1, 1].boxplot(data['Cloud'])
axs[1, 1].set_xlabel('Cloud')

axs[1, 2].boxplot(data['Pressure'])
axs[1, 2].set_xlabel('Pressure')

axs[1, 3].boxplot(data['Wind Speed'])
axs[1, 3].set_xlabel('Wind Speed')

plt.show()
```

HÌNH 1. 7 DỮ LIỆU CỦA BOXPLOT

Sau khi em gọi các cột dữ liệu Weather thì cung cấp cái nhìn tổng quan về phân phối và sự biến động của các yếu tố khí hậu trong tập dữ liệu thời tiết. Mỗi biểu đồ đại diện cho một thuộc tính khác nhau, bao gồm nhiệt độ (Temp), cảm giác (Feels), tốc độ gió (Gust), lượng mưa (Rain), độ ẩm (Humidity), tình trạng mây (Cloud), áp suất (Presure), và tốc độ (Wind Speed)



HÌNH 1. 8 BẢNG CỘT BOXPLOT

Boxplot là một công cụ trực quan hữu ích trong phân tích dữ liệu, cho phép người dùng nhanh chóng nhận diện các đặc điểm chính của phân phối dữ liệu, bao gồm giá trị trung vị, các phần tư, và các điểm ngoại lệ. Qua đó, hình ảnh này không chỉ giúp đánh giá sự phân bố của từng yếu tố khí hậu mà còn chỉ ra những điểm bất thường có thể ảnh hưởng đến các phân tích và mô hình dự đoán sau này. Việc hiểu rõ các yếu tố này là rất quan trọng trong việc đưa ra những quyết định chính xác dựa trên dữ liệu thời tiết.

Biểu đồ này không chỉ giúp người phân tích hiểu rõ hơn về sự phân bố của từng biến mà còn hỗ trợ trong việc phát hiện các vấn đề tiềm ẩn trong dữ liệu, từ đó đưa ra các quyết định chính xác hơn trong quá trình phân tích và mô hình hóa.

4.3.2 Phân bố không phải dạng số

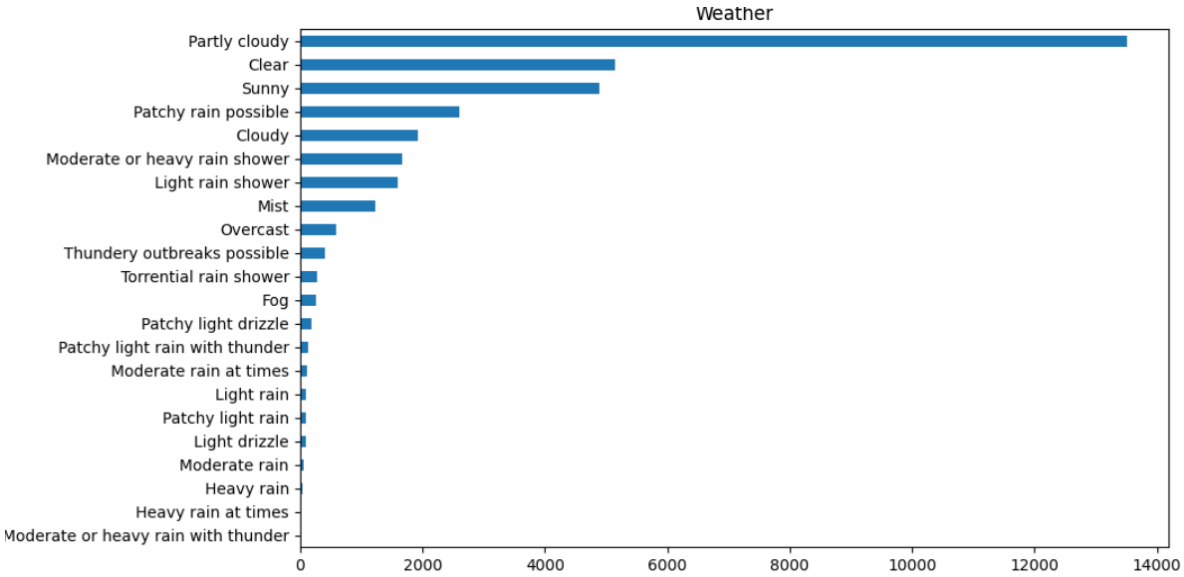
- **Weather**

Qua những thống kê và trực quan tần suất của từng loại thời tiết trong bộ dữ liệu data thông qua các dòng lệnh hàm dic_weather để tạo một dictionary trống để lưu trữ số lượng xuất hiện của từng thời tiết, để duyệt qua các loại thời thời duy nhất. Từ lệnh for weather in ra data nhưng trong đó có vòng lặp Weather

tính số hàng dòng tương ứng với từng loại thời tiết cụ thể , còn giá trị còn được lưu vào dic_weather với key là loại thời tiết và value là số lần xuất hiện .

Khi em tạo một series từ ditionary và sắp xếp theo tần xuất :

- Dictonary dic_weather được sắp xếp theo tăng dần xuất hiện
- Chuyển đổi dictonary đã sắp xếp thành một pandas.Seriea để dễ dàng thao tác biểu đồ



HÌNH 1. 9 BIỂU ĐỒ THANH NGANG (HORIZONTAL BAR CHART)

Sau khi kết quả biểu đồ thanh ngang thì em gọi lệnh value_counts() để đếm xem hàm pandas dùng để đếm số lần xuất hiện bao nhiêu lần của từng giá trị duy nhất :

```
Weather
Partly cloudy                          13513
Clear                                   5139
Sunny                                   4887
Patchy rain possible                    2594
Cloudy                                  1928
Moderate or heavy rain shower           1676
Light rain shower                       1603
Mist                                    1229
Overcast                                 596
Thundery outbreaks possible              413
Torrential rain shower                   283
Fog                                      266
Patchy light drizzle                     180
Patchy light rain with thunder           130
Moderate rain at times                   119
Light rain                               101
Patchy light rain                         99
Light drizzle                             90
Moderate rain                             65
Heavy rain                                33
Heavy rain at times                       30
Moderate or heavy rain with thunder        2
Name: count, dtype: int64
```

HÌNH 1. 10 BẢNG THỐNG KÊ SỐ LƯỢNG THỜI TIẾT

Bảng dữ liệu cung cấp thông tin thống kê về tần xuất hiện của trạng thái thời tiết trong thời tiết trong một khoảng thời gian nhất định. Mỗi hàng trong bảng đại diện cho một loại thời tiết cụ thể, kèm them số lượng lần xuất hiện của nó trong tập dữ liệu. Partly cloudy xuất hiện 13,513 lần trong khi “ clear ” có 5,139 lần và “ Sunny” xuất hiện 4,487. Các trạng thái thời tiết sẽ xuất hiện lần lượt số lần nhiều hơn và giúp phân tích tần suất của trạng thái thời tiết, từ đó cung cấp cái nhìn sâu sắc về điều kiện thời gian trong khu vực và thời gian.

+ Sau đó em gộp các giá trị Weather thành các giá trị Sunny, Clear, Cloudy,Mist và Overcast thanh những biến và em dùng các lệnh để biến các giá

```
array(['Mist', 'Rain', 'Cloudy', 'Sunny', 'Clear', 'Overcast'],
      dtype=object)
```

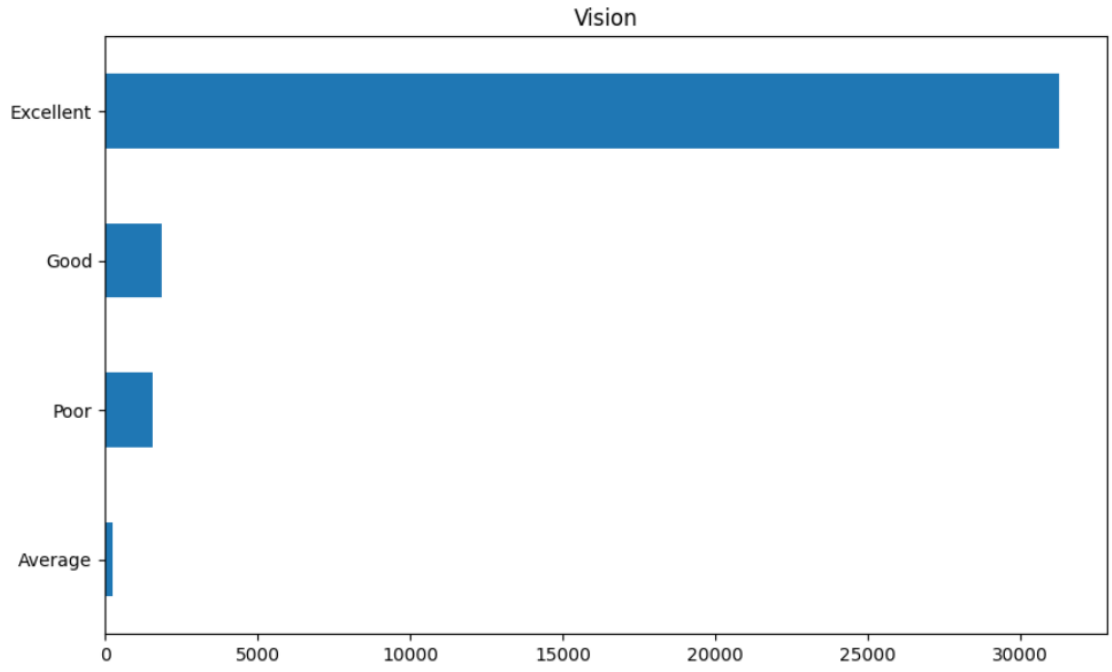
HÌNH 1. 11 GỘP CÁC LỆNH GIÁ TRỊ

trị của convert_weather thì tương ứng với phần tử trong cột Weather) và thực hiện các điều kiện tạo ra kết quả

Kết quả trên là một mảng chứa các trạng thái thời tiết đa dạng, phản ánh sự phong phú của điều kiện khí hậu. Các trạng thái như "Mist" (sương mù) và "Rain" (mưa) gợi nhớ đến những ngày ẩm ướt, trong khi "Sunny" (nắng) và "Clear" (trời quang) mang lại cảm giác tươi sáng và ấm áp. "Cloudy" (nhiều mây) và "Overcast" (trời âm u) lại tạo nên bầu không khí trầm lắng, thường gắn liền với những cơn mưa bất chợt , và các giá trị được gộp chung thành array.

+ Tầm nhìn

Khi em tạo một dictionary để lưu trữ tần xuất từng giá trị trong cột Thì hàm data['Vis'].unique() trả hàm tất cả giá trị duy nhất (không trùng lặp) trong cột (tầm nhìn) và vòng lặp qua từng giá trị duy nhất của cột Vis. data[data['Vis'] == vis].shape[0] thì lọc dữ liệu để lấy tất cả các dòng giá trị Vis bằng vis và đếm số dòng bằng .shape đạt được



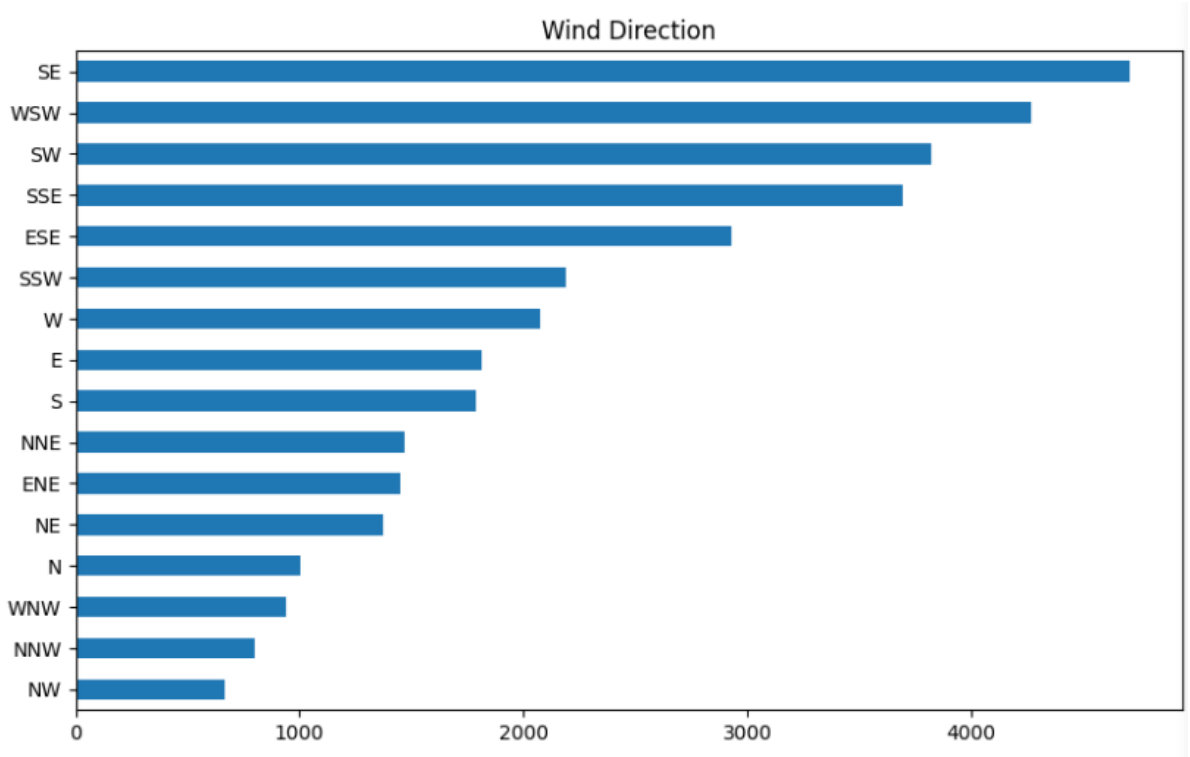
HÌNH 1. 12 BIỂU ĐỒ NGANG VISION

Ta thấy bảng biểu đồ nang vis thì 1 ngang của Excellent dài hơn và các hàng giảm dần và hàm Average là ngắn nhất trong từng biểu đồ thanh . Các thanh trên biểu đồ thể hiện tần suất xuất hiện của từng mức tầm nhìn, được sắp xếp

tăng dần từ tần suất thấp nhất đến tần suất cao nhất. Điều này giúp người xem dễ dàng nhận ra các mức tầm nhìn nào xuất hiện thường xuyên và mức nào ít gặp hơn. Qua từng biểu đồ , có thể thấy rằng một số mức tầm nhìn chiếm tỷ lệ lớn hơn hẳn, biểu thị cho các điều kiện thời tiết phổ biến trong khoảng thời gian thu thập dữ liệu. Những thanh dài đại diện cho các mức tầm nhìn xuất hiện thường xuyên, cho thấy điều kiện quan sát ổn định, rõ ràng trong phần lớn thời gian. Ngược lại, các thanh ngắn hơn thể hiện các mức tầm nhìn hạn chế, có thể liên quan đến các hiện tượng thời tiết bất lợi như sương mù, mưa lớn hoặc khói bụi, làm giảm khả năng quan sát.

+ Win Direction

Em thêm cột ngang của dữ liệu win direction tất cả các cột để duy trì hàm và sau đó còng lặp for qua từng mỗi hướng gió các vòng dữ liệu có hướng gió bằng Win_dir và đếm số lần xuất hiện theo hướng gió và có để theo thanh ngang và đặt kích thước của biểu đồ (chiều rộng 10 , chiều cao 6) .



HÌNH 1. 13 BIỂU ĐỒ THANH NGANG WIND DIRETION

Kết quả đạt được hiển thị các hướng gió hướng gió trong bộ dữ liệu. Trục dọc của biểu đồ hiển thị các hướng gió cụ thể , còn trục ngang biểu diễn số lần

xuất hiện tương ứng của từng hướng gió. Các thanh trong biểu đồ được sắp xếp tăng dần từ hướng gió có tần suất nhỏ nhất đến hướng gió có tần suất lớn nhất, giúp người xem dễ dàng quan sát và so sánh các giá trị. Hướng gió nào càng phổ biến thì thanh biểu diễn của nó càng dài, ngược lại, các hướng gió ít xuất hiện sẽ có thanh ngắn hơn. Việc sắp xếp này mang lại cái nhìn trực quan hơn, giúp nhận ra ngay xu hướng phân bố của các hướng gió trong dữ liệu, từ đó có thể đưa ra các phân tích sâu hơn như tìm hiểu sự thay đổi của hướng gió theo mùa, thời gian hoặc khu vực địa lý."

4.3.3 Phân bố thuộc tính

Sau khi em gọi các biểu đồ và em kiểm tra dữ liệu thêm một lần nữa file csv:

	Date	Time	Weather	Temp	Feels	Gust	Rain	Humidity	Cloud	Pressure	Vis	Wind Speed	Wind Direction
0	Thu 01, Jan 2009	00:00	Mist	23.0	25.0	15.0	0.0	0.97	1.00	1010.0	Poor	9.0	NNW
1	Thu 01, Jan 2009	03:00	Rain	22.0	25.0	13.0	0.4	0.97	0.84	1010.0	Poor	9.0	NNW
2	Thu 01, Jan 2009	06:00	Mist	22.0	25.0	8.0	0.0	0.98	1.00	1011.0	Poor	6.0	N
3	Thu 01, Jan 2009	09:00	Cloudy	27.0	31.0	7.0	0.1	0.83	0.64	1011.0	Excellent	6.0	NNE
4	Thu 01, Jan 2009	12:00	Cloudy	28.0	34.0	3.0	0.0	0.76	0.62	1010.0	Excellent	3.0	NE
5	Thu 01, Jan 2009	15:00	Rain	27.0	32.0	4.0	3.1	0.83	0.74	1009.0	Good	2.0	NNE
6	Thu 01, Jan 2009	18:00	Cloudy	24.0	27.0	14.0	0.0	0.91	0.73	1010.0	Excellent	7.0	NW
7	Thu 01, Jan 2009	21:00	Rain	23.0	26.0	18.0	1.6	0.91	0.32	1012.0	Excellent	10.0	NNE
8	Fri 02, Jan 2009	00:00	Cloudy	22.0	22.0	14.0	0.0	0.91	0.30	1011.0	Excellent	8.0	N
9	Fri 02, Jan 2009	03:00	Cloudy	21.0	22.0	16.0	0.0	0.91	0.72	1010.0	Excellent	9.0	N

HÌNH 1. 14 ĐỌC DỮ LIỆU ĐÃ TIỀN XỬ LÝ

-Thì bộ dữ liệu và thêm hai cột bảng Wind Speed và Wind Direction .

Sau đó bảng dữ liệu này cung cấp cái nhìn sâu sắc về các yếu tố khí hậu trong một khoảng thời gian nhất định, cho phép người phân tích theo dõi và đánh giá các điều kiện thời tiết. Các giá trị nhiệt độ và độ ẩm được ghi lại một cách chi tiết, giúp người dùng hiểu rõ hơn về cảm giác thực tế mà con người trải nghiệm trong điều kiện thời tiết cụ thể.

	Time	Temp	Weather_1	Weather_2	Weather_3	Temp_1	Temp_2	Temp_3	Feels_1	Feels_2
0	09:00	27.0	Mist	Rain	Mist	22.0	22.0	23.0	25.0	25.0
1	12:00	28.0	Cloudy	Mist	Rain	27.0	22.0	22.0	31.0	25.0
2	15:00	27.0	Cloudy	Cloudy	Mist	28.0	27.0	22.0	34.0	31.0
3	18:00	24.0	Rain	Cloudy	Cloudy	27.0	28.0	27.0	32.0	34.0
4	21:00	23.0	Cloudy	Rain	Cloudy	24.0	27.0	28.0	27.0	32.0

HÌNH 1. 15 GIÁ TRỊ CATEGORICAL

Nhìn chung bảng dữ liệu thời tiết trên cung cấp một cái nhìn tổng quan về các điều kiện khí hậu trong một ngày cụ thể, với thông tin chi tiết về thời gian, nhiệt độ và các yếu tố thời tiết khác nhau. Mỗi hàng trong bảng đại diện cho một thời điểm cụ thể trong ngày, cho phép người dùng theo dõi sự thay đổi của thời tiết theo từng giờ.

Nhiệt độ được ghi lại trong cột "Temp" cho thấy sự biến động của nhiệt độ trong suốt cả ngày, từ mức cao nhất là 28.0°C vào buổi trưa đến mức thấp nhất là 23.0°C vào buổi tối. Điều này cho thấy sự thay đổi nhiệt độ tự nhiên theo thời gian, thường thấy trong các khu vực có khí hậu nhiệt đới. cột "Weather_1", "Weather_2" và "Weather_3" cung cấp thông tin về các điều kiện thời tiết khác nhau, cho thấy rằng thời tiết có thể rất đa dạng trong cùng một ngày. Ví dụ, vào buổi sáng, thời tiết có thể có sương mù, trong khi vào buổi chiều có thể có mưa và đến tối lại trở nên nhiều mây.

Điều này cho thấy sự phức tạp của khí hậu và tầm quan trọng của việc theo dõi các yếu tố thời tiết để đưa ra các dự đoán chính xác. Cuối cùng, các cột "Temp_1", "Temp_2" và "Temp_3" cho phép so sánh nhiệt độ giữa các nguồn khác nhau, giúp người dùng hiểu rõ hơn về sự khác biệt trong các phép đo nhiệt độ. Việc này có thể hữu ích trong việc đánh giá độ chính xác của các thiết bị đo lường hoặc trong việc phân tích các điều kiện khí hậu tại các vị trí khác nhau.

+ Tiếp đến em kiểm tra các độ tương quan giữa các hàm giữa thuộc tính của hàm Temp

	Temp
Humidity_1	-0.593766
Temp_3	-0.201505
Gust_2	-0.200011
Feels_3	-0.185120
Gust_3	-0.161530
Humidity_2	-0.154154
Cloud_2	-0.142889
Gust_1	-0.123309
Cloud_1	-0.121738
Rain_3	-0.119025
Wind Speed_3	-0.113688
Cloud_3	-0.109521
Rain_2	-0.106488
Rain_1	-0.082713
Wind Speed_2	-0.045941
Pressure_1	0.026635
Wind Speed_1	0.094662
Pressure_2	0.124769
Pressure_3	0.133886
Humidity_3	0.192592
Temp_2	0.198147
Feels_2	0.216687
Feels_1	0.701572
Temp_1	0.720200

HÌNH 1. 16 THUỘC TÍNH TEMP

Hình tương quan trên thể hiện mối quan hệ giữa thuộc tính "Temp" (Nhiệt độ) và các thuộc tính khác trong tập dữ liệu. Các giá trị tương quan nằm trong khoảng từ -1 đến 1, cho biết mức độ liên hệ giữa các biến:

+ Tương quan âm

- Humidity_1 (-0.593766): Độ ẩm ở thời điểm trước đó có mối tương quan tiêu cực mạnh với nhiệt độ, cho thấy rằng khi độ ẩm tăng, nhiệt độ có xu hướng giảm.
- Temp_3 (-0.201505): Nhiệt độ ở hai thời điểm trước đó cũng có mối tương quan tiêu cực yếu với nhiệt độ hiện tại.

- Các thuộc tính khác như Gust_2 (-0.200011), Feels_3 (-0.185120), và Gust_3 (-0.161530) cũng cho thấy mối tương quan tiêu cực, nhưng ở mức độ yếu hơn.

+ Tương quan dương

- Temp_1 (0.720200): Nhiệt độ ở thời điểm trước đó có mối tương quan mạnh với nhiệt độ hiện tại, cho thấy rằng nhiệt độ trước đó có ảnh hưởng lớn đến nhiệt độ hiện tại.
- Feels_1 (0.701572): Cảm giác nhiệt độ ở thời điểm trước đó cũng có mối tương quan mạnh với nhiệt độ hiện tại.
- Các thuộc tính khác như Temp_2 (0.198147) và Humidity_3 (0.192592) có mối tương quan dương yếu với nhiệt độ.

+ Tương quan gần bằng 0

Kết luận: Những thuộc tính có mối tương quan cao với "Temp" như Temp_1 và Feels_1 có thể được sử dụng làm đặc trưng (features) trong mô hình dự đoán nhiệt độ, trong khi những thuộc tính có mối tương quan thấp có thể được loại bỏ để giảm độ phức tạp của mô hình.

+ Thuộc tính categarical cho việc dự đoán

```
data.drop(columns=['Wind Direction_1', 'Wind Direction_2', 'Wind Direction_3', 'Vis_1', 'Vis_2', 'Vis_3'], inplace=True)
data.head()
```

	Time	Temp	Weather_1	Weather_2	Weather_3	Temp_1	Feels_1	Humidity_1
0	09:00	27.0	Mist	Rain	Mist	22.0	25.0	0.98
1	12:00	28.0	Cloudy	Mist	Rain	27.0	31.0	0.83
2	15:00	27.0	Cloudy	Cloudy	Mist	28.0	34.0	0.76
3	18:00	24.0	Rain	Cloudy	Cloudy	27.0	32.0	0.83
4	21:00	23.0	Cloudy	Rain	Cloudy	24.0	27.0	0.91

HÌNH 1. 17 THUỘC TÍNH CATEGARICAL

Dữ liệu trong bảng trên thể hiện các thông số khí tượng học tại nhiều thời điểm khác nhau trong ngày, giúp ta có cái nhìn chi tiết về các điều kiện thời tiết. Đại diện cho một thời điểm cụ thể, như 09:00, 12:00, 15:00, và đi kèm với các giá trị quan trọng như nhiệt độ, thời tiết, độ ẩm và cảm giác nhiệt độ thực tế. Cột "Time" biểu thị thời gian ghi nhận dữ liệu, tạo thành một chuỗi thời gian theo từng khung giờ cụ thể. "Temp" phản ánh nhiệt độ thực tế được đo tại thời điểm đó, trong khi "Feels_1" cho thấy nhiệt độ mà con người cảm nhận được, thường bị ảnh hưởng bởi các yếu tố như độ ẩm hoặc gió. Bảng dữ liệu này là kết quả của việc làm sạch dữ liệu ban đầu, trong đó các cột không cần thiết như hướng gió (Wind Direction_1, Wind Direction_2, Wind Direction_3) và tầm nhìn (Vis_1, Vis_2, Vis_3) đã được loại bỏ. Việc này giúp cho dữ liệu trở nên rõ ràng, dễ hiểu và tập trung vào các yếu tố quan trọng nhất.

Nhìn tổng thể, dữ liệu này đóng vai trò quan trọng trong việc phân tích các xu hướng thời tiết theo thời gian, phục vụ cho các mục đích như dự báo thời tiết, đánh giá ảnh hưởng của khí hậu đến đời sống con người hoặc nghiên cứu các hiện tượng thời tiết đặc biệt. Bảng dữ liệu không chỉ cung cấp thông tin một cách đầy đủ mà còn tạo tiền đề để thực hiện các phân tích sâu hơn về mối quan hệ giữa nhiệt độ, độ ẩm và hiện tượng thời tiết.

+ Các loại giá trị phân loại hot encoding

	Temp	Temp_1	Feels_1	Humidity_1	Time_00:00	Time_03:00	Time_06:00	Time_09:00	Time_12:00	Time_15:00
0	27.0	22.0	25.0	0.98	False	False	False	True	False	False
1	28.0	27.0	31.0	0.83	False	False	False	False	True	False
2	27.0	28.0	34.0	0.76	False	False	False	False	False	True
3	24.0	27.0	32.0	0.83	False	False	False	False	False	False
4	23.0	24.0	27.0	0.91	False	False	False	False	False	False

HÌNH 1. 18 PHÂN LOẠI GIÁ TRỊ HOT ENCODING

- Temp: Nhiệt độ hiện tại (đơn vị có thể là độ C).
- Temp_1: Nhiệt độ cảm nhận (có thể là cảm giác nhiệt độ).
- Feels_1: Nhiệt độ cảm nhận thực tế.
- Humidity_1: Độ ẩm không khí, thể hiện dưới dạng số thập phân (0-1).

- Time_00:00, Time_03:00, Time_06:00, Time_09:00: Các cột này cho biết trạng thái thời tiết tại các thời điểm cụ thể trong ngày, với giá trị True hoặc False để chỉ ra sự hiện diện của một điều kiện thời tiết nhất định.

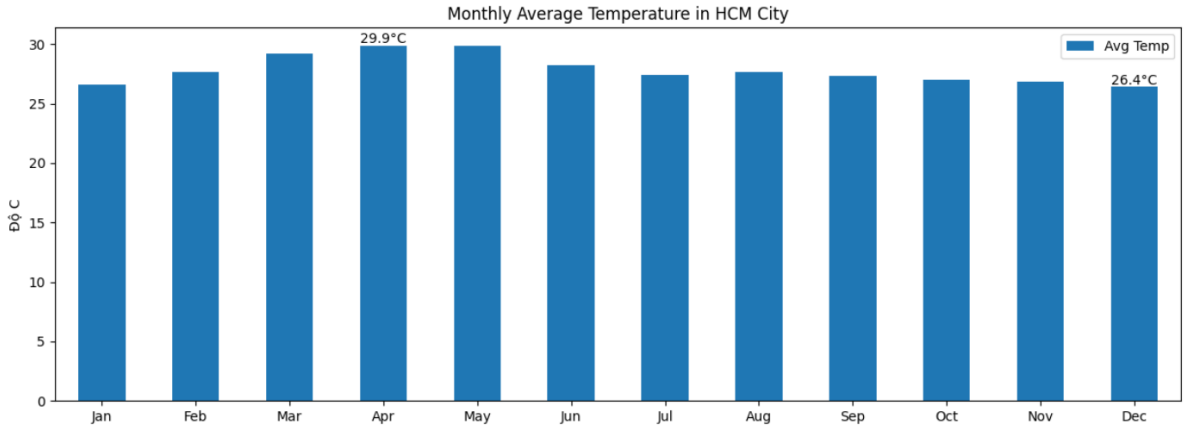
+ Bảng dữ liệu này cung cấp cái nhìn sâu sắc về các yếu tố khí hậu trong một khoảng thời gian nhất định, cho phép người phân tích theo dõi và đánh giá các điều kiện thời tiết. Các giá trị nhiệt độ và độ ẩm được ghi lại một cách chi tiết, giúp người dùng hiểu rõ hơn về cảm giác thực tế mà con người trải nghiệm trong điều kiện thời tiết cụ thể.

Cột temp cho ta thấy nhiệt độ hiện tại, trong khi Feels_1 phản ánh cảm giác mà người dùng có thể cảm nhận được, có thể bị ảnh hưởng bởi độ ẩm và gió. Độ ẩm được thể hiện qua cột Humidity_1, cho thấy mức độ ẩm trong không khí, điều này rất quan trọng trong việc đánh giá sự thoải mái của con người.

Các cột thời gian cho phép theo dõi sự thay đổi của các yếu tố khí hậu theo thời gian, với các giá trị True và False cho biết liệu một điều kiện thời tiết cụ thể có xảy ra hay không tại các thời điểm khác nhau trong ngày. Điều này không chỉ giúp người dùng có cái nhìn tổng quan về thời tiết mà còn hỗ trợ trong việc đưa ra các quyết định liên quan đến hoạt động ngoài trời, nông nghiệp, và các lĩnh vực khác liên quan đến khí hậu.

4.3.4 Phân bố nhiệt độ trung bình

Tiếp đến chúng ta kiểm Dữ liệu thời tiết, việc hiểu rõ các tình trạng thời tiết là rất quan trọng để đưa ra các dự đoán chính xác. Biểu đồ dưới đây thể hiện phân bố của các tình trạng thời tiết trong dữ liệu thu thập được từ thành phố Hồ Chí Minh. Thông qua biểu đồ này, chúng ta có thể nhận thấy các tình trạng thời tiết phổ biến nhất và tần suất xuất hiện của chúng trong khoảng thời gian từ ngày 01/01/2009 đến ngày 20/12/2020 .



HÌNH 1. 19 HÌNH ẢNH TRUNG BÌNH NHIỆT ĐỘ THEO THÁNG

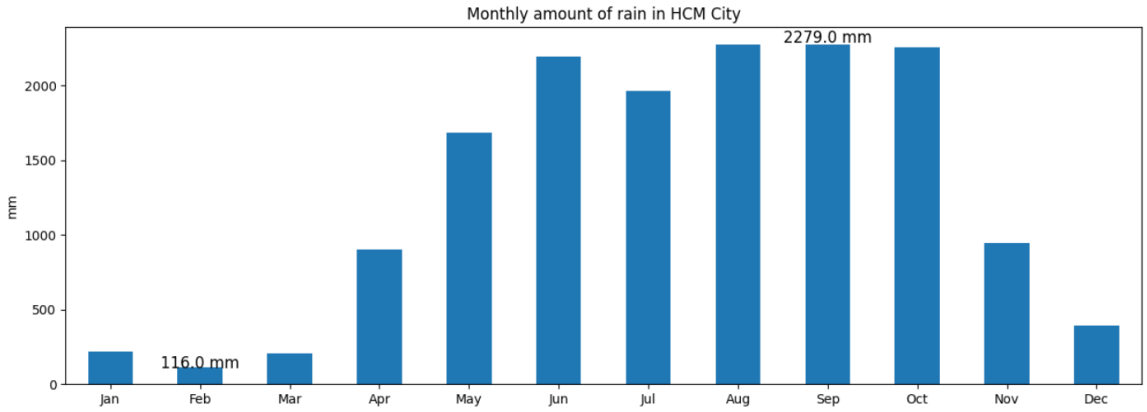
- Ta thấy sự nhiệt độ trung bình theo biểu đồ cho thấy nhiệt độ trung bình hàng tháng ở Hồ Chí Minh dao động từ khoảng 25°C đến 30°C. Nhiệt độ cao nhất thường rơi vào tháng 4 (April), cho thấy đây là thời điểm nóng nhất trong năm.Nhiệt độ trung bình: Nhiệt độ trung bình hàng tháng dao động từ khoảng 25°C đến 30°C, với một số tháng có nhiệt độ cao hơn, nhưng không có tháng nào dưới 25°C.

Thay vào đó sự ổn định sự biến động nhỏ, nhiệt độ ở Hồ Chí Minh nhìn chung khá ổn định và không có sự thay đổi lớn giữa các tháng. Điều này phản ánh khí hậu nhiệt đới của thành phố, nơi mà thời tiết thường ấm áp và ít biến động.

Ý nghĩa về nhiệt động trung bình hàng tháng rất quan trọng cho việc lập kế hoạch nông nghiệp, du lịch và các hoạt động ngoài trời. Nó cũng giúp người dân và các nhà quản lý có cái nhìn tổng quan về điều kiện khí hậu trong khu vực , biểu đồ nhiệt trên cung cấp cái nhìn tổng quan về điều kiện khí hậu tại Hồ Chí Minh, cho thấy sự ổn định và tính chất nhiệt đới của khu vực. Thông tin này rất hữu ích cho việc phân tích và dự đoán thời tiết trong tương lai

4.3.5 Phân bố lượng mưa trung bình tháng

Dữ liệu thời tiết, lượng mưa là một yếu tố quan trọng trong khí hậu của một khu vực, ảnh hưởng đến nông nghiệp, sinh hoạt và các hoạt động ngoài trời. Biểu đồ dưới đây thể hiện lượng mưa trung bình hàng tháng tại thành phố Hồ Chí Minh. Thông qua biểu đồ này, chúng ta có thể nhận thấy các tháng có lượng mưa cao nhất và thấp nhất, từ đó hiểu rõ hơn về xu hướng khí hậu của khu vực.



HÌNH 1. 20 LƯỢNG MƯA TRUNG BÌNH THEO THÁNG

+ Lượng mưa hàng tháng: Biểu đồ cho thấy lượng mưa trung bình hàng tháng ở Hồ Chí Minh có sự biến động lớn. Các tháng như June, July, August, và September có lượng mưa cao, với một số tháng đạt trên 2000 mm. Điều này cho thấy mùa mưa ở thành phố thường rơi vào khoảng thời gian này.

+ Tháng 2 (February): Lượng mưa của tháng 2 là thấp nhất trong năm, gần như không đáng kể. Điều này phản ánh rằng tháng 2 thường là một phần của mùa khô ở Hồ Chí Minh.

+ Xu hướng mùa mưa: Biểu đồ rõ ràng cho thấy xu hướng mùa mưa và mùa khô của thành phố, điều này rất quan trọng cho việc lập kế hoạch nông nghiệp và các hoạt động ngoài trời.

Biểu đồ lượng mưa hàng tháng cung cấp cái nhìn tổng quan về điều kiện khí hậu tại Hồ Chí Minh, cho thấy sự biến động rõ rệt giữa mùa mưa và mùa khô. Thông tin này rất hữu ích cho việc phân tích và dự đoán thời tiết trong tương lai.

4.4 Phân tích kết quả

4.4.1 Đánh giá mô hình

Dựa trên các chỉ số trên, chúng ta có thể đánh giá mô hình hồi quy tuyến tính như sau:

```
(0.9389432582425412,
 21.466159974767226,
 array([[32.01393738]]),
 array([-44.16854198]))
```

HÌNH 1. 21 ĐÁNH GIÁ MÔ HÌNH

Độ phù hợp của mô hình:

Với hệ số R^2 cao (0.93), mô hình có khả năng dự đoán tốt và giải thích được phần lớn sự biến thiên của dữ liệu. Điều này cho thấy mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc là rõ ràng và đáng tin cậy.

Sai số dự đoán:

Mặc dù có sai số trung bình tuyệt đối (MAE = 21,46), mức sai số này vẫn chấp nhận được, đặc biệt khi dữ liệu đầu vào có nhiễu hoặc tính biến động tự nhiên.

Hệ số:

Hệ số hồi quy (32.01) cho thấy mức độ ảnh hưởng mạnh mẽ của biến độc lập đến biến phụ thuộc. Hệ số chặn (-44,16) cho biết giá trị YYY khi X=0X = 0X=0.

4.4.2 Kết quả mô hình

Tập dữ liệu mô hình mô hình Polynomial Regression

HÌNH 1. 22 TẬP DỮ LIỆU MÔ HÌNH

+ Kết luận về hiệu số mô hình

Dựa vào hai chỉ số R^2 và MAE, chúng ta có thể kết luận rằng mô hình hồi quy tuyến tính hiện tại hoạt động rất tốt. Cụ thể:

$R^2 = 0.99$ cho thấy mô hình giải thích được hầu hết sự biến thiên trong dữ liệu, thể hiện mối quan hệ tuyến tính mạnh mẽ giữa các biến.

MAE = 6,98 là mức sai số thấp, cho thấy khả năng dự đoán chính xác của mô hình.

- Mô hình rất phù hợp với dữ liệu: Giá trị R^2 gần đạt mức tối đa (1.0), cho thấy rằng mô hình hầu như không để sót các thông tin quan trọng trong dữ liệu.
- Độ chính xác cao: MAE thấp chứng tỏ sai lệch giữa dự đoán của mô hình và thực tế rất nhỏ. Điều này đặc biệt quan trọng trong các bài toán cần tính chính xác, chẳng hạn như dự báo tài chính, y tế, hay khoa học dữ liệu.

+ **Hạn chế cần cân nhắc:**

Mặc dù kết quả rất khả quan, cần lưu ý rằng:

- Mức R^2 cao có thể là dấu hiệu của hiện tượng overfitting nếu dữ liệu kiểm tra chưa được sử dụng để đánh giá. Do đó, cần kiểm tra mô hình trên một tập dữ liệu độc lập (validation set) để xác nhận hiệu quả thực tế.
- Giá trị MAE chỉ phản ánh trung bình sai số tuyệt đối, do đó không cung cấp thông tin về phân phối sai số (ví dụ, sai số lớn ở một số điểm ngoại lệ).

+ Kết quả vẽ của mô hình

HÌNH 1. 23 KẾT QUẢ ĐẠT ĐƯỢC

- Sự tương đồng: Hai đường biểu diễn có xu hướng di chuyển song song, cho thấy rằng mô hình đã dự đoán khá chính xác nhiệt độ trong khoảng thời gian này.
- Biến động: Cả hai đường đều có sự biến động, phản ánh sự thay đổi nhiệt độ theo thời gian. Mô hình có thể đã nắm bắt được các xu hướng chính trong dữ liệu.
- Sai số: Mặc dù có sự tương đồng, vẫn có một số điểm mà đường dự đoán không hoàn toàn khớp với đường thực tế. Điều này có thể chỉ ra rằng mô hình có thể cải thiện hơn nữa để giảm sai số.

Biểu đồ này cung cấp cái nhìn trực quan về hiệu suất của mô hình hồi quy tuyến tính trong việc dự đoán nhiệt độ. Sự tương đồng giữa giá trị dự đoán và giá trị thực tế cho thấy mô hình hoạt động tốt, nhưng vẫn cần tiếp tục cải thiện để đạt được độ chính xác cao hơn

CHƯƠNG V: KẾT LUẬN VÀ KIẾN NGHỊ

5.1 Hạn chế của đề tài

+ Khả năng quá khớp (Overfitting)

Dự án phân tích dữ liệu khí tượng, mặc dù có nhiều tiềm năng và ứng dụng thực tiễn, nhưng cũng gặp phải một số hạn chế đáng lưu ý. Đầu tiên, chất lượng dữ liệu là một yếu tố quan trọng; dữ liệu khí tượng có thể bị thiếu hoặc không chính xác do lỗi trong quá trình thu thập, dẫn đến kết quả phân tích không đáng tin cậy. Hơn nữa, nếu dữ liệu chỉ được thu thập từ một khu vực cụ thể, nó có thể không phản ánh được các điều kiện khí hậu rộng hơn, làm giảm tính tổng quát của các kết quả.

Thêm vào đó, các mô hình dự đoán có thể không chính xác nếu không được tối ưu hóa hoặc nếu không có đủ dữ liệu để huấn luyện, điều này có thể dẫn đến những dự đoán sai lệch về thời tiết. Các yếu tố bên ngoài như biến đổi khí hậu, thiên tai, và các hiện tượng thời tiết cực đoan cũng có thể ảnh hưởng đến dữ liệu và làm cho các phân tích trở nên phức tạp hơn. Các bên liên quan cũng là một yếu tố quan trọng; nếu dự án không có sự tương tác này, có thể dẫn đến việc không đáp ứng được nhu cầu thực tế của cộng đồng. Bên cạnh đó, chi phí và tài nguyên cần thiết cho việc thu thập và phân tích dữ liệu khí tượng có thể đòi hỏi một khoản đầu tư lớn, điều này có thể hạn chế khả năng thực hiện dự án.

Cuối cùng, việc trực quan hóa dữ liệu khí tượng phức tạp có thể gặp khó khăn, đặc biệt khi cần thể hiện nhiều yếu tố cùng một lúc. Thời tiết có thể thay đổi nhanh chóng, và các mô hình dự đoán có thể không kịp thời phản ánh những thay đổi này, dẫn đến thông tin không còn chính xác. Những hạn chế này cần được xem xét và giải quyết trong quá trình thực hiện dự án để đảm bảo rằng kết quả đạt được là chính xác và có giá trị cho người dùng.

5.2 Hướng phát triển

Thử nghiệm với các mô hình phức tạp hơn: mô hình nâng cao hơn và đảm bảo tính tổng quát của mô hình trong các tình huống thực tế. Bên cạnh đó, việc triển khai mô hình vào thực tế thông qua các ứng dụng công nghệ sẽ giúp mô hình trở nên hữu ích và có giá trị cao.

Những nỗ lực này không chỉ giúp nâng cao độ chính xác của mô hình mà còn mở rộng khả năng ứng dụng của nó trong các lĩnh vực khác nhau như kinh tế, giáo dục, nông nghiệp, và công nghiệp. Việc không ngừng cải tiến và phát triển mô hình sẽ tạo ra một công cụ mạnh mẽ hỗ trợ dự đoán và ra quyết định trong tương lai. Kết nối với các API thời tiết để cung cấp dữ liệu thời gian thực, giúp người dùng có thông tin cập nhật và chính xác hơn.

Cải thiện giao diện người dùng: Thiết kế lại giao diện ứng dụng để thân thiện hơn với người dùng, bao gồm các biểu đồ tương tác và các tùy chọn lọc dữ liệu nâng cao. Phân tích nâng cao: Thêm các phân tích thống kê và mô hình hóa dữ liệu để người dùng có thể hiểu rõ hơn về các yếu tố ảnh hưởng đến thời tiết.Hỗ trợ nhiều định dạng dữ liệu: Mở rộng khả năng của ứng dụng để hỗ trợ nhiều định dạng tệp khác nhau, không chỉ giới hạn ở CSV.

TÀI LIỆU KHAM KHẢO

1. Sách và Tài liệu Chuyên ngành:

- "Meteorology: Understanding the Atmosphere" - Steven A. Ackerman, John A. Knox

- "Data Analysis Methods in Physical Oceanography" - William J. Emery, Richard E. Thomson

2. Bài báo Khoa học:

- "A review of data assimilation techniques in meteorology" - P. D. Williams, et al.

- "Statistical Methods for Weather and Climate" - Daniel Wilks

3. Tài liệu Hướng dẫn và Tài nguyên Trực tuyến:

- Pandas Documentation: Hướng dẫn sử dụng thư viện pandas cho phân tích dữ liệu.

- Matplotlib Documentation: Tài liệu hướng dẫn về cách trực quan hóa dữ liệu bằng Matplotlib.

4. Cơ sở Dữ liệu Khí tượng:

- NOAA National Centers for Environmental Information: Cung cấp dữ liệu khí tượng và khí hậu từ khắp nơi trên thế giới.

- World Meteorological Organization (WMO): Tổ chức cung cấp thông tin và dữ liệu khí tượng toàn cầu.

5. Khóa học Trực tuyến:

- "Introduction to Meteorology" trên Coursera hoặc edX: Các khóa học cung cấp kiến thức cơ bản về khí tượng học.

- "Data Science for Everyone" trên DataCamp: Khóa học giúp hiểu rõ hơn về phân tích dữ liệu.