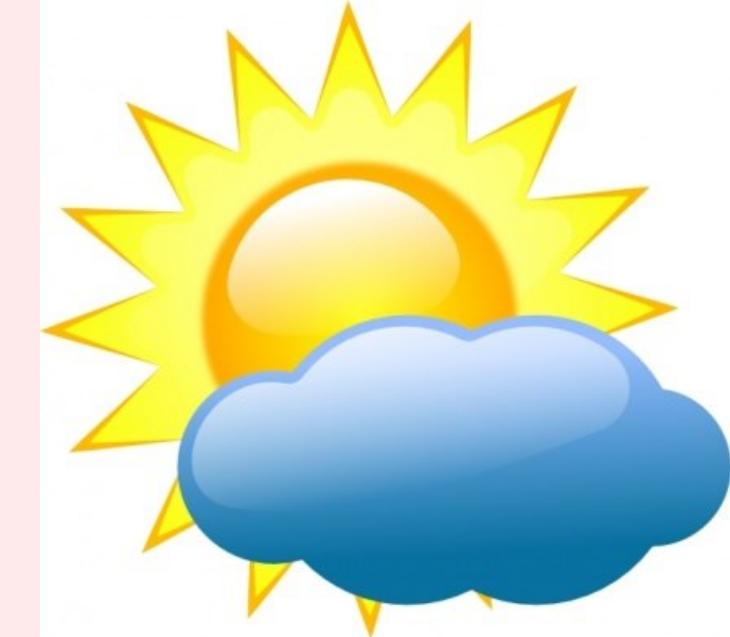


**TRƯỜNG ĐẠI HỌC KINH TẾ
KHOA THỐNG KÊ – TIN HỌC**



BÁO CÁO THỰC TẬP NGHỀ NGHIỆP

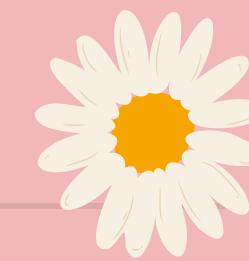
**ÁP DỤNG MÔ HÌNH HỌC MÁY TRONG PHÂN TÍCH VÀ DỰ ĐOÁN DỮ LIỆU
THỜI TIẾT**



**Sinh viên thực hiện :Nguyễn Thị Trà My
Lớp : 46K21.1
Giảng viên hướng dẫn : TS. Phan Đình Văn**



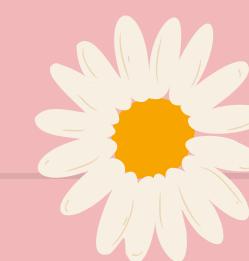
CHƯƠNG I : TỔNG QUAN VỀ ĐỀ TÀI VÀ GIỚI THIỆU ĐƠN VỊ THỰC TẬP



CHƯƠNG 2. CƠ SỞ LÝ THUYẾT PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO THỜI TIẾT



CHƯƠNG 3. PHƯƠNG PHÁP THU THẬP VÀ XỬ LÝ DỮ LIỆU



CHƯƠNG 4 : HỆ THỐNG PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO THỜI TIẾT

CHƯƠNG I : TỔNG QUAN VỀ ĐỀ TÀI VÀ GIỚI THIỆU ĐƠN VỊ THỰC TẬP



1.1.Giới thiệu tổng quát về công ty TNHH phần mềm Việt Đà



1.1.1.Tổng quan về doanh nghiệp

Thứ nhất, thông tin công ty

Tên công ty: Công ty TNHH phần mềm Việt Đà

Tên quốc tế: VIETDA SOFTWARE COMPANY LIMITED

Tên viết tắt: VIETDA SOFTWARE

Mã số thuế: 0400621146

Người đại diện theo pháp luật: Nguyễn Hậu

Điện thoại: 02363726926

Ngày hoạt động: 08/05/2008



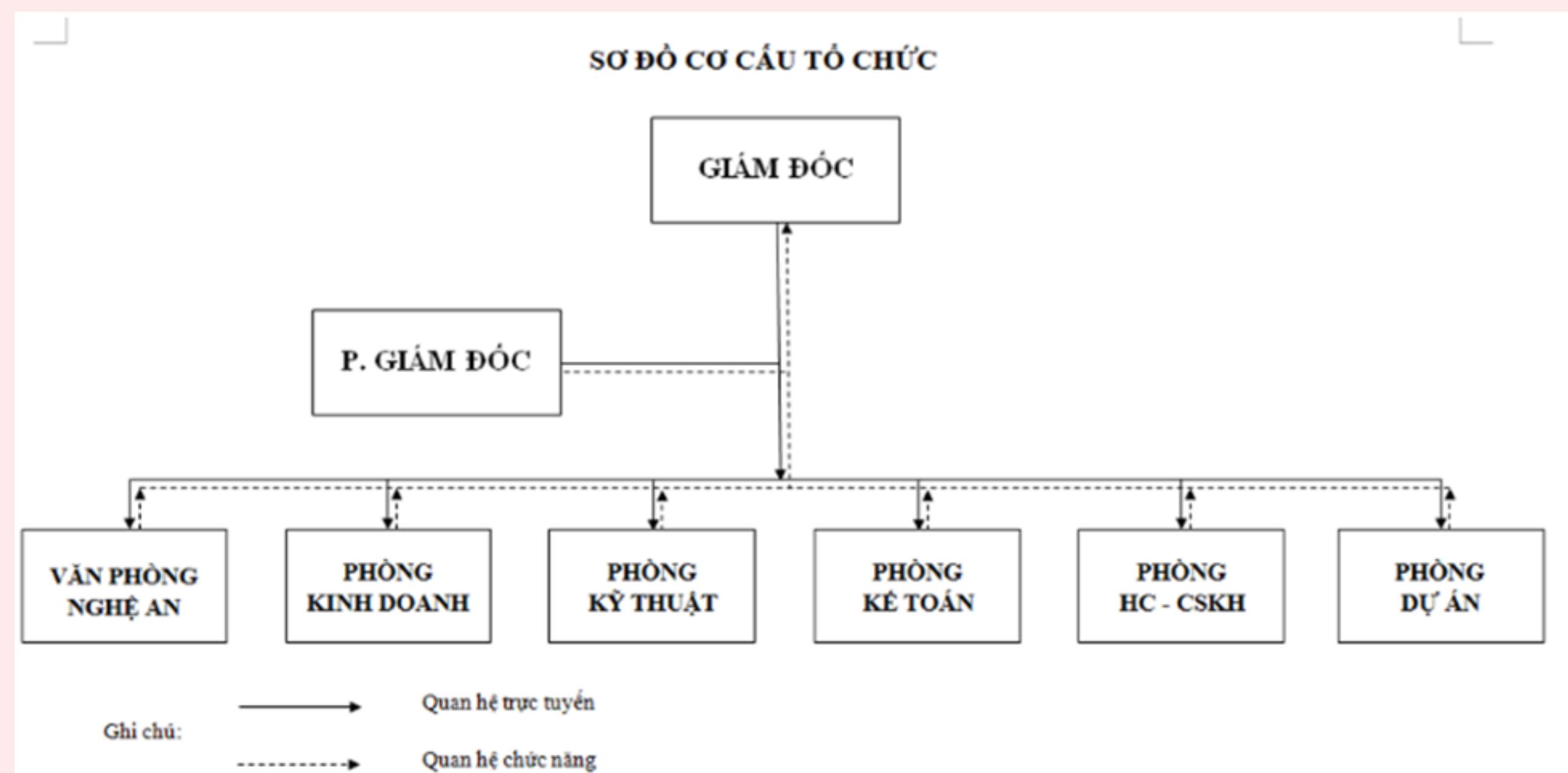
Lịch sử hình thành và phát triển

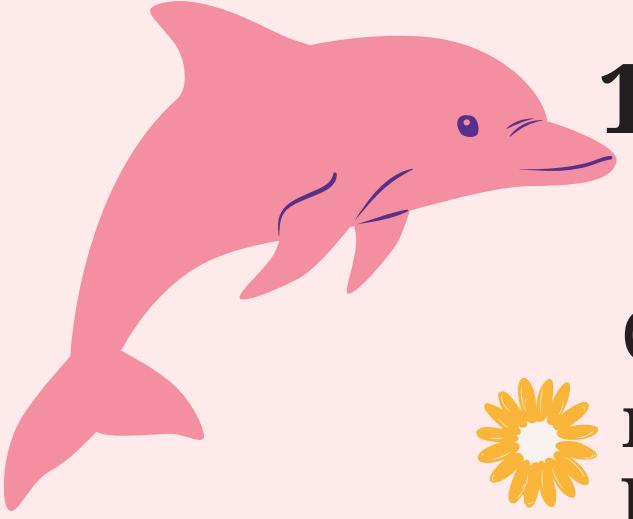
Công ty TNHH Phần Mềm Việt Đà được thành lập ngày 30/4/2008 với chức năng chính là sản xuất và cung ứng phần mềm ứng Ngày 28/3/2014, thành lập Văn Phòng Đại Diện Nghệ An tại Tp. Vinh, Tỉnh Nghệ An nhằm cung cấp phần mềm và hỗ trợ tốt hơn cho khách hàng các tỉnh Bắc Trung Bộ và Bắc Bộ.

Trong khi đó, trụ sở công ty chuyển văn phòng về địa chỉ: 41 Trần Văn Ông, P. Hòa An, Q.Cẩm Lệ, Tp.Đà Nẵng.

Hiện nay, Phần Mềm Việt Đà đã phân phối phần mềm đến tất cả 63 tỉnh thành trong cả nước. Đáp ứng đa dạng nhu cầu của khách hàng theo từng nhóm ngành hoặc đặc thù riêng của doanh nghiệp và các quy định của cơ quan quản lý.

Sơ đồ cơ cấu tổ chức





1.1.2.Tầm nhìn, sứ mệnh và ngành nghề kinh doanh của công ty



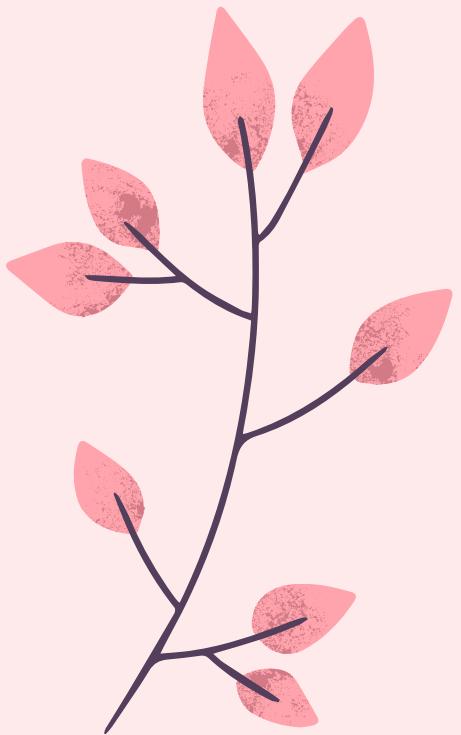
Công ty TNHH Phần mềm Việt Đà xác định tầm nhìn trở thành công ty phần mềm hàng đầu Việt Nam, cung cấp các giải pháp và sản phẩm phần mềm có chất lượng cao, đáp ứng mọi nhu cầu quản lý và vận hành doanh nghiệp của khách hàng.

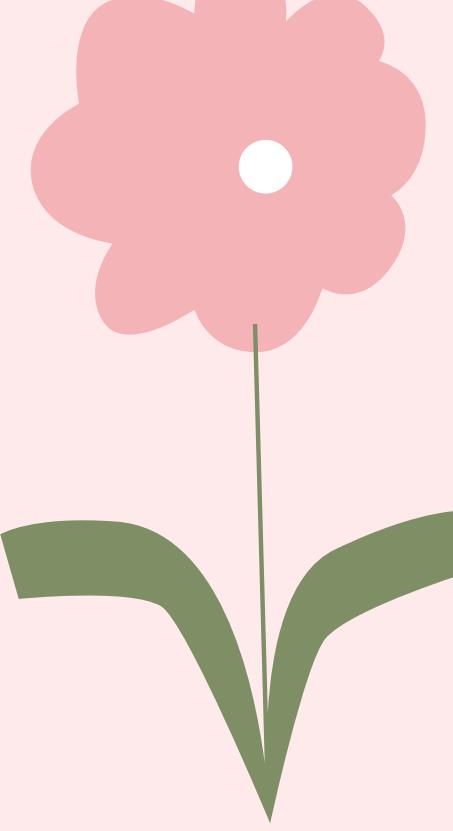


Công ty TNHH Phần mềm Việt Đà xác định sứ mệnh của mình là trở thành cầu nối giữa công nghệ và kinh doanh, mang đến cho khách hàng những giải pháp CNTT hiệu quả và tiện ích nhất, góp phần nâng cao năng lực quản trị và năng suất kinh doanh cho các tổ chức, doanh nghiệp tại Việt Nam.



Công ty TNHH Phần mềm Việt Đà hoạt động trong lĩnh vực cung cấp các giải pháp phần mềm ứng dụng cho doanh nghiệp. Cụ thể, các sản phẩm và dịch vụ chính của Công ty bao gồm: Phần mềm kế toán doanh nghiệp; phần mềm nhà hàng, khách sạn; phần mềm quản lý bán hàng; phần mềm quản lý kho, hàng hóa và phần mềm quản trị nhân lực.





1.2.Tổng quan về đề tài

1.2.1.Vấn đề thực tiễn

Vấn đề thực tiễn về việc áp dụng mô hình học máy trong phân tích và dự đoán dữ liệu thời tiết đang trở thành một lĩnh vực nghiên cứu quan trọng và hứa hẹn. Trong thời đại số hóa và thông tin hiện nay, khả năng dự báo chính xác về thời tiết có vai trò quan trọng trong nhiều lĩnh vực, từ an toàn công cộng đến nông nghiệp và công nghiệp năng lượng.

Mô hình học máy là một phương pháp dự đoán dựa trên việc học từ dữ liệu và tự điều chỉnh để cải thiện hiệu suất dự đoán

Tuy nhiên, việc áp dụng mô hình học máy trong phân tích và dự đoán dữ liệu thời tiết cũng đặt ra một số vấn đề thực tiễn cần được xem xét: Dữ liệu không chắc chắn và không đồng nhất; độ phức tạp của mô hình; cơ sở dữ liệu lớn và xử lý dữ liệu; điều kiện biên và biến đổi thời tiết cực đoan;...



1.2.2 Định hướng phát triển đề tài

Định hướng phát triển đề tài "Áp dụng mô hình học máy trong phân tích và dự đoán dữ liệu thời tiết" là một hướng nghiên cứu hứa hẹn với tiềm năng cải thiện đáng kể khả năng dự báo thời tiết và ứng dụng thực tiễn trong nhiều lĩnh vực. Dưới đây là một số hướng mở rộng và phát triển tiềm năng cho đề tài này:

- Tối ưu mô hình học máy
- Kết hợp dữ liệu đa nguồn
- Xử lý dữ liệu thời tiết không chắc chắn
- Dự báo thời tiết cực đoan
- Phân tích ứng dụng thực tế
- Khám phá phân tích dữ liệu sâu hơn

Trong tương lai, việc phát triển và áp dụng mô hình học máy trong phân tích và dự đoán dữ liệu thời tiết sẽ đóng góp quan trọng vào việc nâng cao khả năng dự báo, bảo vệ an toàn và cải thiện hiệu suất của nhiều hoạt động quan trọng.



CHƯƠNG 2. CƠ SỞ LÝ THUYẾT PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO THỜI TIẾT

2.1. Giới thiệu về Data Analyst

- Vị trí công việc Data Analyst đang trở nên ngày càng quan trọng và phổ biến trong xã hội hiện nay. Sự phát triển nhanh chóng của công nghệ thông tin và khai phá dữ liệu đã tạo ra một lượng lớn thông tin và dữ liệu có giá trị. Data Analyst đóng vai trò quan trọng trong việc biến các dữ liệu này thành thông tin hữu ích và ý nghĩa cho các doanh nghiệp, tổ chức, và xã hội.
- Data Analyst giúp các doanh nghiệp và tổ chức nắm bắt được những xu hướng, thông tin và insights quan trọng từ dữ liệu để đưa ra quyết định thông minh và hiệu quả.
- Chính phủ sử dụng Data Analyst để tối ưu hóa các chính sách và dự đoán xu hướng xã hội. Trong lĩnh vực y tế, Data Analyst giúp cải thiện dịch vụ chăm sóc sức khỏe và dự đoán các bệnh lý. Trong giáo dục, họ đóng vai trò quan trọng trong phân tích dữ liệu về học sinh và đánh giá hiệu quả giáo dục. Nhu cầu về Data Analyst đang ngày càng tăng cao, và đây là một trong những ngành nghề đang có nhu cầu cao về nhân lực.

- Data Analyst là một chuyên gia trong lĩnh vực phân tích dữ liệu, có nhiệm vụ thu thập, xử lý, phân tích và hiểu dữ liệu từ các nguồn khác nhau nhằm tìm ra các thông tin, xu hướng và quan trọng.
- Data Analyst có khả năng làm việc với các nguồn dữ liệu lớn và phức tạp, từ dữ liệu số liệu doanh thu, thông tin khách hàng, tới dữ liệu về sản phẩm và dịch vụ.
- Data Analyst áp dụng các phương pháp phân tích thống kê, học máy và khai phá dữ liệu để khám phá sâu hơn thông tin tiềm ẩn trong dữ liệu.
- Với sự phát triển mạnh mẽ của khoa học dữ liệu và xu hướng số hóa, Data Analyst đã trở thành một trong những vị trí công việc có nhu cầu cao và hấp dẫn trong thời đại hiện đại.

2.1.2. Công việc của Data Analyst trong doanh nghiệp:

- Công việc của Data Analyst trong doanh nghiệp là tập trung vào việc phân tích và hiểu rõ dữ liệu để hỗ trợ quyết định kinh doanh và cải thiện hiệu suất hoạt động. Dưới đây là một số nhiệm vụ chính mà Data Analyst thực hiện trong môi trường doanh nghiệp:
 - Thu thập dữ liệu
 - Xử lý và làm sạch dữ liệu
 - Phân tích dữ liệu
 - Đưa ra insights và giải pháp
 - Theo dõi và đánh giá hiệu suất
 - Dự báo và kế hoạch tương lai
- Tóm lại, Data Analyst đóng vai trò quan trọng trong việc chuyển đổi dữ liệu thành thông tin hữu ích và ý nghĩa cho doanh nghiệp. Công việc của họ giúp cải thiện quyết định kinh doanh, tối ưu hóa hoạt động và đảm bảo hiệu suất kinh doanh cao hơn.

2.2. Tổng quan về dữ liệu thời tiết và kinh doanh

- Dữ liệu thời tiết đóng vai trò quan trọng trong hoạt động kinh doanh, đó là những thông tin về tình trạng thời tiết như nhiệt độ, lượng mưa, tốc độ gió, độ ẩm, áp suất không khí và nhiều yếu tố khác trong một khu vực cụ thể trong một khoảng thời gian nhất định. Thông tin này được thu thập từ các trạm quan trắc, vệ tinh, cảm biến và các nguồn dữ liệu khác.

- Đối với hoạt động kinh doanh, dữ liệu thời tiết đóng góp một số lợi ích quan trọng trong:

- + Chiến lược tiếp thị
- + Quản lý chuỗi cung ứng
- + Quản lý rủi ro
- + Dự báo năng suất nông nghiệp

- Tóm lại, dữ liệu thời tiết đóng vai trò quan trọng trong hoạt động kinh doanh, giúp các doanh nghiệp hiểu rõ hơn về môi trường kinh doanh và đưa ra các quyết định thông minh và hiệu quả.



66

2.2.1 Khái niệm về dữ liệu thời tiết:



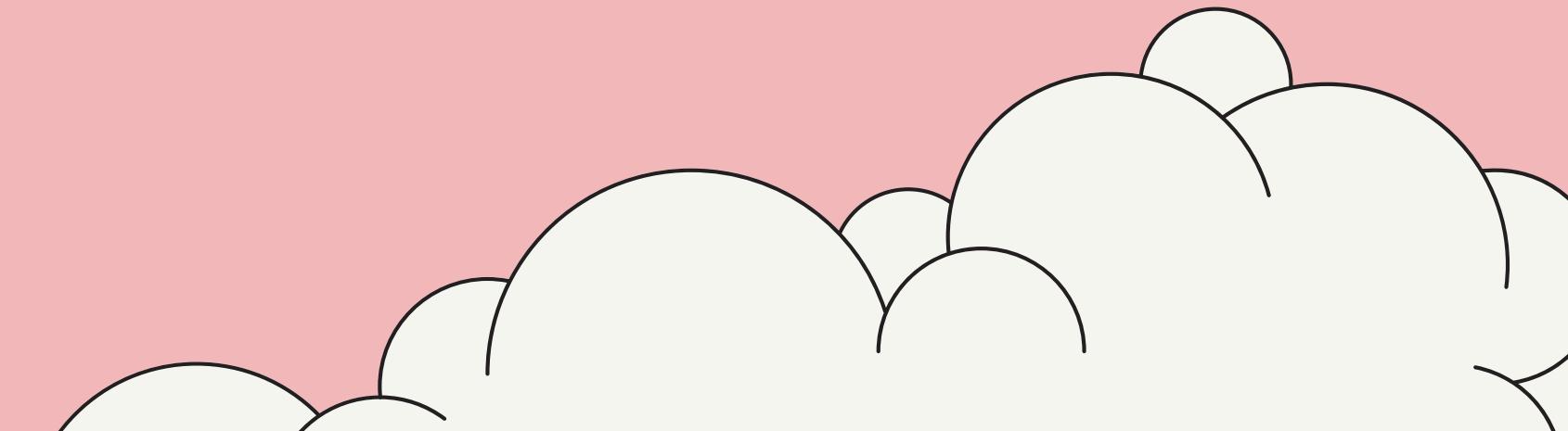
Dữ liệu thời tiết là tập hợp các thông tin về các điều kiện và hiện tượng thời tiết tại một vị trí cụ thể trong một khoảng thời gian nhất định. Thông tin này bao gồm các yếu tố như nhiệt độ, lượng mưa, tốc độ gió, độ ẩm, áp suất không khí, thời tiết tổng quan và nhiều thông số khác.



Dữ liệu thời tiết đóng vai trò quan trọng trong nhiều lĩnh vực như dự báo thời tiết, quản lý thiên tai, nông nghiệp, giao thông vận tải, ngành công nghiệp năng lượng, du lịch và các ngành công nghiệp liên quan đến thời tiết. Dự báo thời tiết giúp người dân và các tổ chức chuẩn bị và ứng phó với các điều kiện thời tiết tiềm ẩn.



Dữ liệu thời tiết cũng hỗ trợ trong việc nghiên cứu các xu hướng và biến đổi khí hậu, đồng thời đánh giá tác động của thời tiết đối với các hoạt động và quyết định kinh doanh.



2.2.2 Tầm quan trọng của dữ liệu thời tiết trong kinh doanh:

- Dữ liệu thời tiết có tầm quan trọng vô cùng trong kinh doanh và đóng vai trò quan trọng trong nhiều khía cạnh của hoạt động kinh doanh. Dưới đây là một số điểm nổi bật về tầm quan trọng của dữ liệu thời tiết trong kinh doanh:
 - Chiến lược tiếp thị và bán hàng
 - Quản lý chuỗi cung ứng và sản xuất
 - Quản lý rủi ro và đối phó với thiên tai
 - Dự báo năng suất nông nghiệp
 - Định hướng kế hoạch và phát triển



2.2.3 Lợi ích và ứng dụng của phân tích dữ liệu thời tiết trong kinh doanh:

Phân tích dữ liệu thời tiết trong kinh doanh mang lại nhiều lợi ích và ứng dụng quan trọng, giúp doanh nghiệp hiểu rõ hơn về điều kiện thời tiết và tối ưu hóa hoạt động kinh doanh. Dưới đây là một số lợi ích và ứng dụng chính của phân tích dữ liệu thời tiết trong kinh doanh:

- Dự báo và quản lý nhu cầu sản phẩm
- Tối ưu hóa chuỗi cung ứng
- Quản lý rủi ro và đối phó thiên tai
- Tối ưu hóa tiêu thụ năng lượng và tài nguyên
- Định hướng quyết định kinh doanh

Phân tích dữ liệu thời tiết trong kinh doanh mang lại nhiều lợi ích và ứng dụng đa dạng. Việc sử dụng thông tin từ dữ liệu thời tiết giúp cải thiện quyết định kinh doanh, tối ưu hóa hoạt động, và giảm thiểu rủi ro trong môi trường kinh doanh biến đổi.



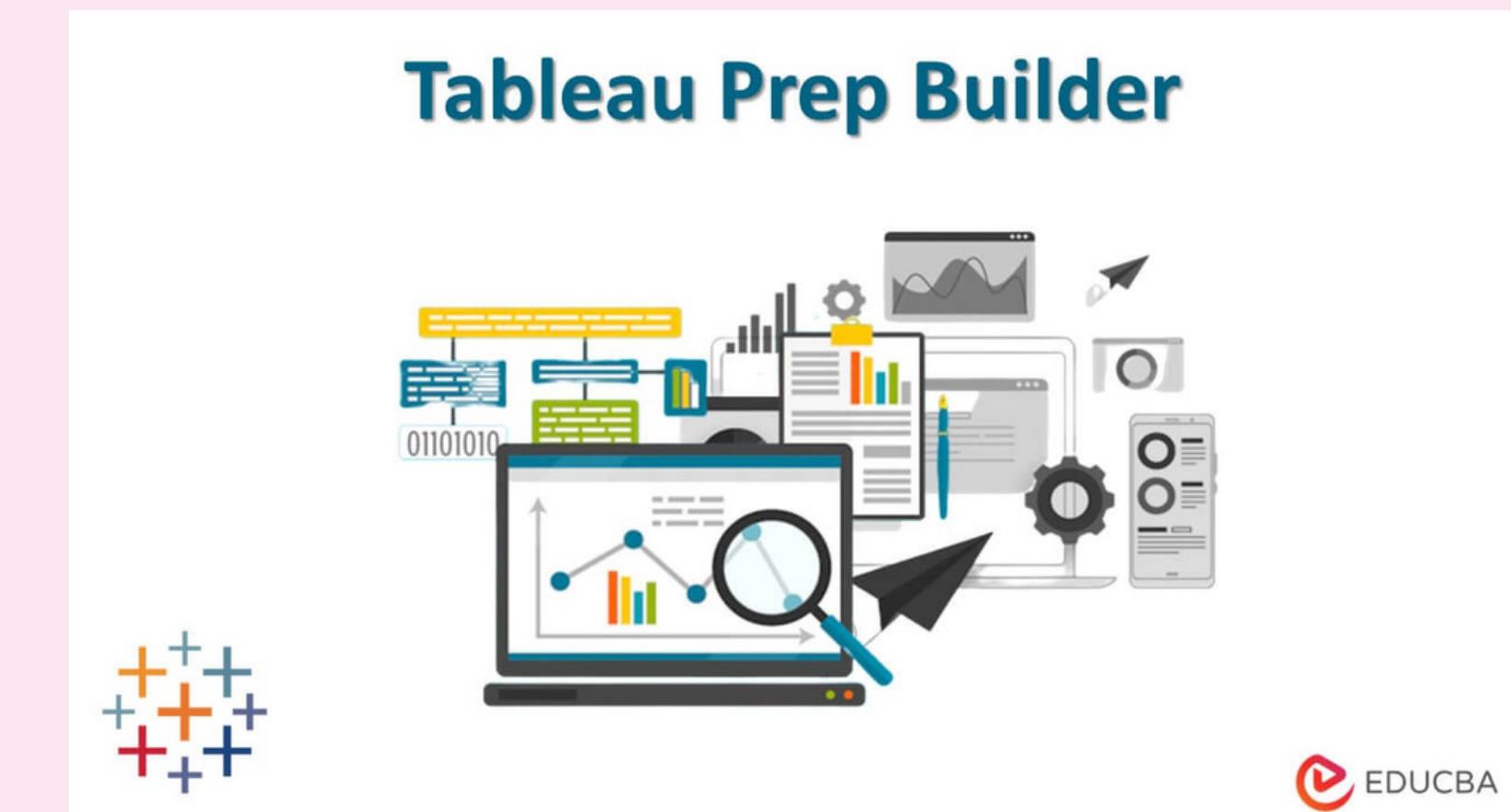
2.3 Công cụ xử lý dữ liệu Tableau Prep Builder

- **Khái niệm:**

Tableau Prep Builder là một công cụ phần mềm do hãng Tableau Software phát triển, được sử dụng để chuẩn bị và làm sạch dữ liệu trước khi phân tích và trực quan hóa bằng phần mềm Tableau. Được ra mắt vào năm 2018, Tableau Prep Builder giúp người dùng kết nối, biến đổi, và làm sạch dữ liệu từ nhiều nguồn khác nhau một cách dễ dàng và hiệu quả.

Với Tableau Prep Builder, người dùng có thể thực hiện các công việc chuẩn bị dữ liệu như:

- Kết nối dữ liệu
- Biến đổi dữ liệu
- Làm sạch dữ liệu
- Tích hợp dữ liệu



2.3.2. Đặc điểm nổi bật Tableau Prep Builder:

- Tableau Prep Builder có một số đặc điểm nổi bật đáng chú ý, giúp người dùng thực hiện công việc chuẩn bị dữ liệu một cách dễ dàng và hiệu quả. Dưới đây là một số đặc điểm nổi bật của Tableau Prep Builder:
 - Giao diện trực quan
 - Chuẩn bị dữ liệu đa nguồn
 - Xem trước dữ liệu
 - Tự động phát hiện sự không phù hợp
 - Lịch sử thay đổi
 - Tích hợp với Tableau Desktop
- Tableau Prep Builder có những đặc điểm nổi bật giúp người dùng thực hiện công việc chuẩn bị dữ liệu một cách trực quan, linh hoạt và hiệu quả. Công cụ này giúp tối ưu hóa quá trình chuẩn bị dữ liệu và đảm bảo tính chính xác của dữ liệu trước khi tiến hành phân tích và trực quan hóa.

2.4 .Công cụ trực quan hóa dữ liệu Tableau

2.4.1. Khái niệm Tableau:

Tableau là một công cụ phân tích dữ liệu và trực quan hóa dữ liệu được phát triển bởi Tableau Software. Nó cho phép người dùng kết nối, thao tác và trực quan hóa dữ liệu từ nhiều nguồn khác nhau một cách dễ dàng và hiệu quả. Tableau được sử dụng rộng rãi trong các lĩnh vực như kinh doanh, khoa học dữ liệu, quản lý, giáo dục và nhiều lĩnh vực khác để giúp người dùng hiểu rõ hơn về dữ liệu và đưa ra quyết định thông minh.



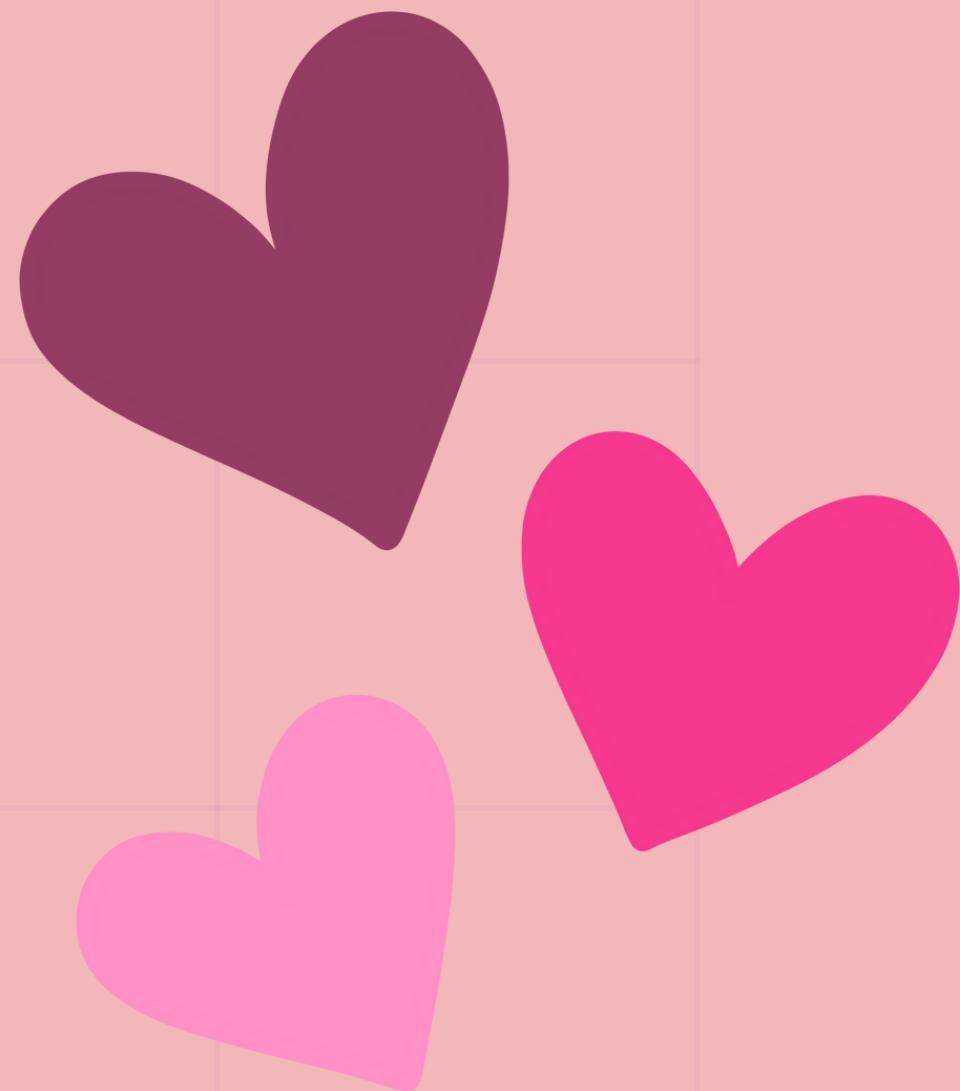
2.4.2. Các tính năng nổi bật của Tableau:

- Trực quan hóa dữ liệu
- Kết nối đa nguồn
- Tích hợp dữ liệu
- Tự động cập nhật
- Khả năng chia sẻ và xuất dữ liệu

Với những tính năng và đặc điểm nổi bật như trên, Tableau đã trở thành một công cụ phân tích dữ liệu mạnh mẽ và phổ biến trong cộng đồng doanh nghiệp và người làm việc với dữ liệu.

2.4.3. Các loại biểu đồ thường dùng:

- 1. Q&A Dashboard**
- 2. Top Down Dashboard**
- 3. Bottom Up Dashboard**
- 4. KPI Dashboard**
- 5. One Big Chart Dashboard**



2.5 Khai phá dữ liệu :

2.5.1 Khái niệm:

Khai phá dữ liệu (Data Mining) là quá trình tìm kiếm thông tin có giá trị, mô hình, xu hướng, hoặc mẫu ẩn chưa biết từ dữ liệu lớn và phức tạp. Nó là một phần của lĩnh vực khoa học dữ liệu và trí tuệ nhân tạo. Mục tiêu của khai phá dữ liệu là khám phá những thông tin hữu ích và tiết lộ sự tương quan, mẫu, hay tri thức từ dữ liệu để hỗ trợ quá trình ra quyết định, dự đoán kết quả tương lai hoặc hiểu sâu về hiện tượng nghiên cứu.

Các bước chính trong quá trình khai phá dữ liệu bao gồm: Thu thập dữ liệu, Tiền xử lý dữ liệu, Chọn phương pháp khai phá, Khai phá dữ liệu, Đánh giá và hiển thị kết quả, Sử dụng kiến thức

2.5.2 Vai trò của việc khai phá dữ liệu:

1. Khám phá thông tin ẩn
2. Dự đoán và dự báo
3. Phát hiện thông tin giá trị
4. Tối ưu hóa quy trình
5. Hiểu biết về khách hàng
6. Phân tích thị trường
7. Phát triển chiến lược kinh doanh
8. Phân loại và nhóm dữ liệu



2.5.3 Các kỹ thuật khai phá dữ liệu:

Có nhiều kỹ thuật khai phá dữ liệu khác nhau, mỗi kỹ thuật phục vụ cho mục tiêu cụ thể trong việc tìm kiếm thông tin ẩn và mẫu trong dữ liệu. Dưới đây là một số kỹ thuật phổ biến trong khai phá dữ liệu:

- Phân tích cụm (Cluster Analysis)
- Phân loại (Classification)
- Hồi quy (Regression)
- Khai thác liên kết (Association Rule Mining)
- Phát hiện bất thường (Anomaly Detection)
- Phân tích chuỗi thời gian (Time Series Analysis)
- Phân tích đồ thị (Graph Analysis)
- Phân tích văn bản (Text Mining)
- Phân tích tương tự (Similarity Analysis)
- Khai phá dữ liệu trực quan (Visual Data Mining)



2.5.4 Các mô hình khai phá dữ liệu:



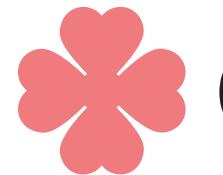
2.5.4.1. Decision Tree – Cây quyết định

Cây quyết định (Decision Tree) là một mô hình học máy được sử dụng trong khai phá dữ liệu và phân loại. Nó biểu diễn dữ liệu dưới dạng cây, trong đó mỗi nút đại diện cho một quyết định hoặc một kiểm tra trên một thuộc tính dữ liệu. Cây quyết định giúp dự đoán giá trị mục tiêu bằng cách tương tác qua các nút và cạnh trên cây từ gốc đến lá.

Mỗi nút quyết định của cây thực hiện một kiểm tra dựa trên một thuộc tính của dữ liệu. Dựa vào kết quả kiểm tra, cây sẽ điều hướng đến một nút con khác, tiếp tục kiểm tra hoặc đưa ra quyết định. Các lá của cây đại diện cho các lớp hoặc nhãn của dữ liệu.



2.5.4 Các mô hình khai phá dữ liệu:



Quá trình xây dựng cây quyết định thường bao gồm các bước sau:

1. Chọn thuộc tính quan trọng
2. Tách dữ liệu
3. Xây dựng nút quyết định
4. Lặp lại các bước trên
5. Cắt tỉa (Pruning)

Cây quyết định có khả năng dễ hiểu và trực quan, giúp trình bày các quyết định và luật quyết định trong dữ liệu một cách rõ ràng. Tuy nhiên, nó có thể dễ dàng bị quá khớp (overfitting) nếu không được kiểm soát cẩn thận.

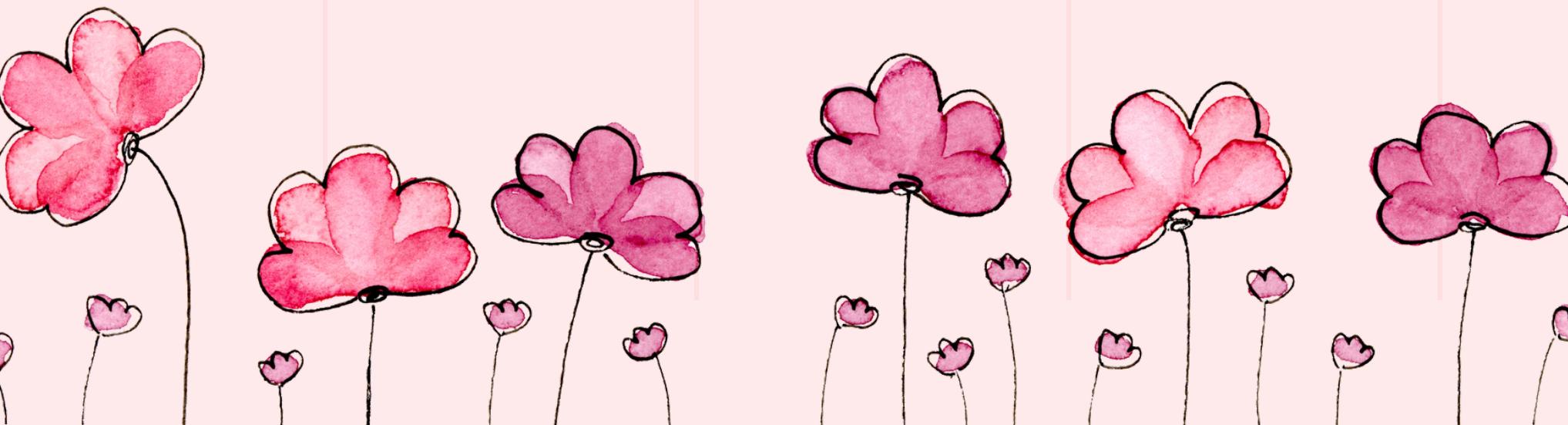
2.5.4.2 . KNN – K-Nearest Neighbor

K-Nearest Neighbors (KNN) là một thuật toán học máy được sử dụng trong phân loại và dự đoán. Ý tưởng cơ bản của KNN là dự đoán lớp hoặc giá trị của một điểm dữ liệu mới bằng cách so sánh nó với các điểm dữ liệu gần nhất (các láng giềng) trong tập dữ liệu đã biết.

Thuật toán KNN hoạt động như sau:

- Chọn số láng giềng (K)
- Tính khoảng cách
- Chọn K láng giềng gần nhất
- Phân loại hoặc dự đoán

KNN dựa vào giả định rằng các điểm dữ liệu cùng lớp sẽ nằm gần nhau trong không gian đặc trưng. Nó là một thuật toán đơn giản và dễ hiểu, thích hợp cho những tập dữ liệu nhỏ hoặc có độ phức tạp thấp. Tuy nhiên, KNN có thể bị ảnh hưởng bởi nhiều dữ liệu và yêu cầu một lượng lớn dữ liệu huấn luyện để đạt được hiệu suất tốt.



2.5.4.3. Random Forest:

Rando Forest là một mô hình học máy được sử dụng cho phân loại dự đoán và khai phá dữ liệu. Nó là một tập hợp (ensemble) của các cây quyết định độc lập, được hình thành bằng cách kết hợp dự đoán của nhiều cây quyết định khác nhau để tạo ra một dự đoán cuối cùng.

Ý tưởng chính của Random Forest là tạo ra nhiều cây quyết định khác nhau bằng cách sử dụng các tập dữ liệu con khác nhau và các đặc trưng ngẫu nhiên Sau đó, khi cần dự đoán, Random Forest kết hợp các dự đoán từ các cây con để tạo ra một dự đoán tổng thể. Quá trình này giúp giảm thiểu tình trạng quá khớp (overfitting) và tạo ra một mô hình ổn định và có khả năng tổng quát hóa tốt.

Một số điểm nổi bật về Random Forest:

1. Phân loại và dự đoán
2. Khả năng xử lý nhiễu
3. Tích hợp các cây quyết định
4. Tính hiệu suất
5. Tránh overfitting

Random Forest đã trở thành một công cụ quan trọng trong học máy và khai phá dữ liệu, được sử dụng trong nhiều ứng dụng khác nhau như dự đoán giá chứng khoán, phát hiện gian lận tín dụng, phân loại ảnh, và nhiều lĩnh vực khác.

2.5.5 Các phương pháp đánh giá mô hình

2.5.5.1 Confusion matrix

1. Confusion Matrix (Ma trận nhầm lẫn) là một công cụ quan trọng trong đánh giá hiệu suất của các mô hình phân loại trong học máy và khai phá dữ liệu. Nó giúp đánh giá sự chính xác của dự đoán bằng cách so sánh giữa dự đoán của mô hình và thực tế.

2. Các khái niệm trong Confusion Matrix:

- True Positive (TP)
- False Positive (FP)
- True Negative (TN)
- False Negative (FN)

2.5.5 Các phương pháp đánh giá mô hình

2.5.5.1 Confusion matrix

1. Confusion Matrix cung cấp thông tin quan trọng để tính toán các chỉ số đánh giá hiệu suất như độ chính xác (accuracy), độ nhạy (sensitivity), độ cụ thể (specificity), dự đoán tích cực (positive predictive value), và dự đoán âm (negative predictive value).
2. Dựa vào Confusion Matrix, bạn có thể đánh giá mô hình của mình và hiểu rõ hơn về các sai sót mà mô hình có thể thực hiện trong việc dự đoán các lớp hoặc nhãn.



2.5.5 Các phương pháp đánh giá mô hình

2.5.5.2. Các chỉ số đánh giá mô hình

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

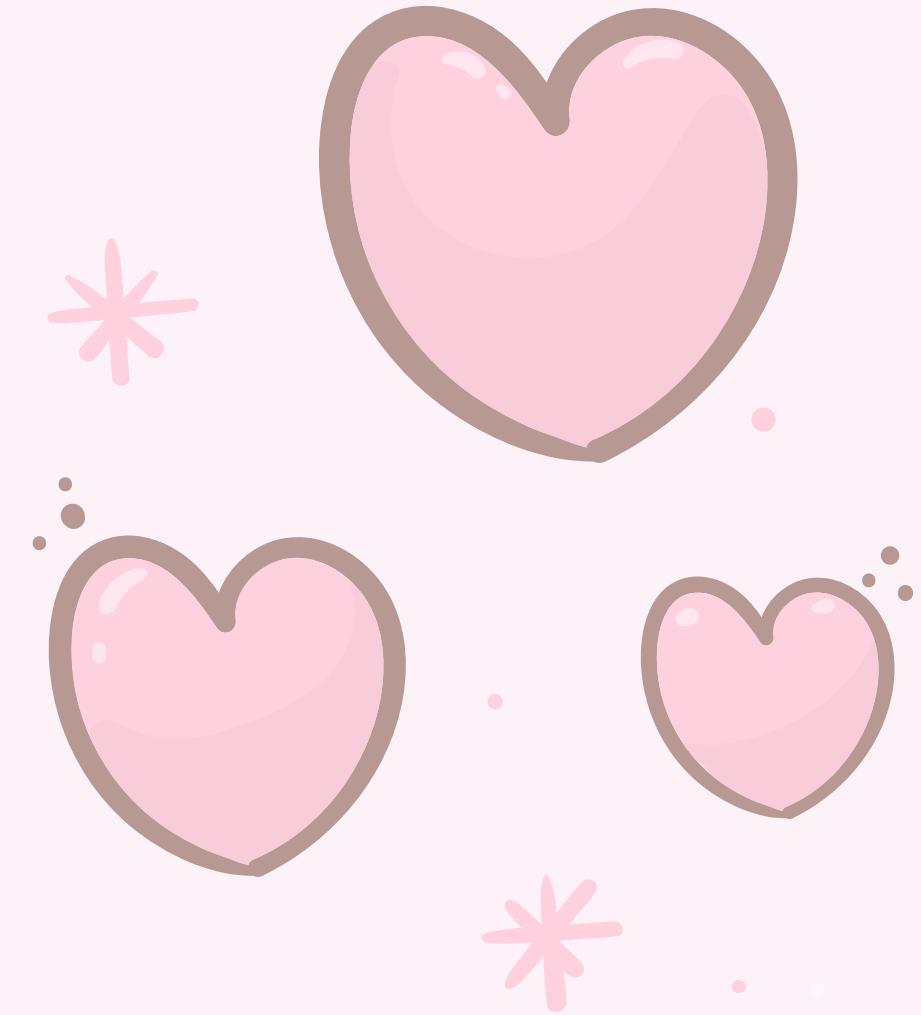
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

2.5.5 Các phương pháp đánh giá mô hình

2.5.5.2. Các chỉ số đánh giá mô hình

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

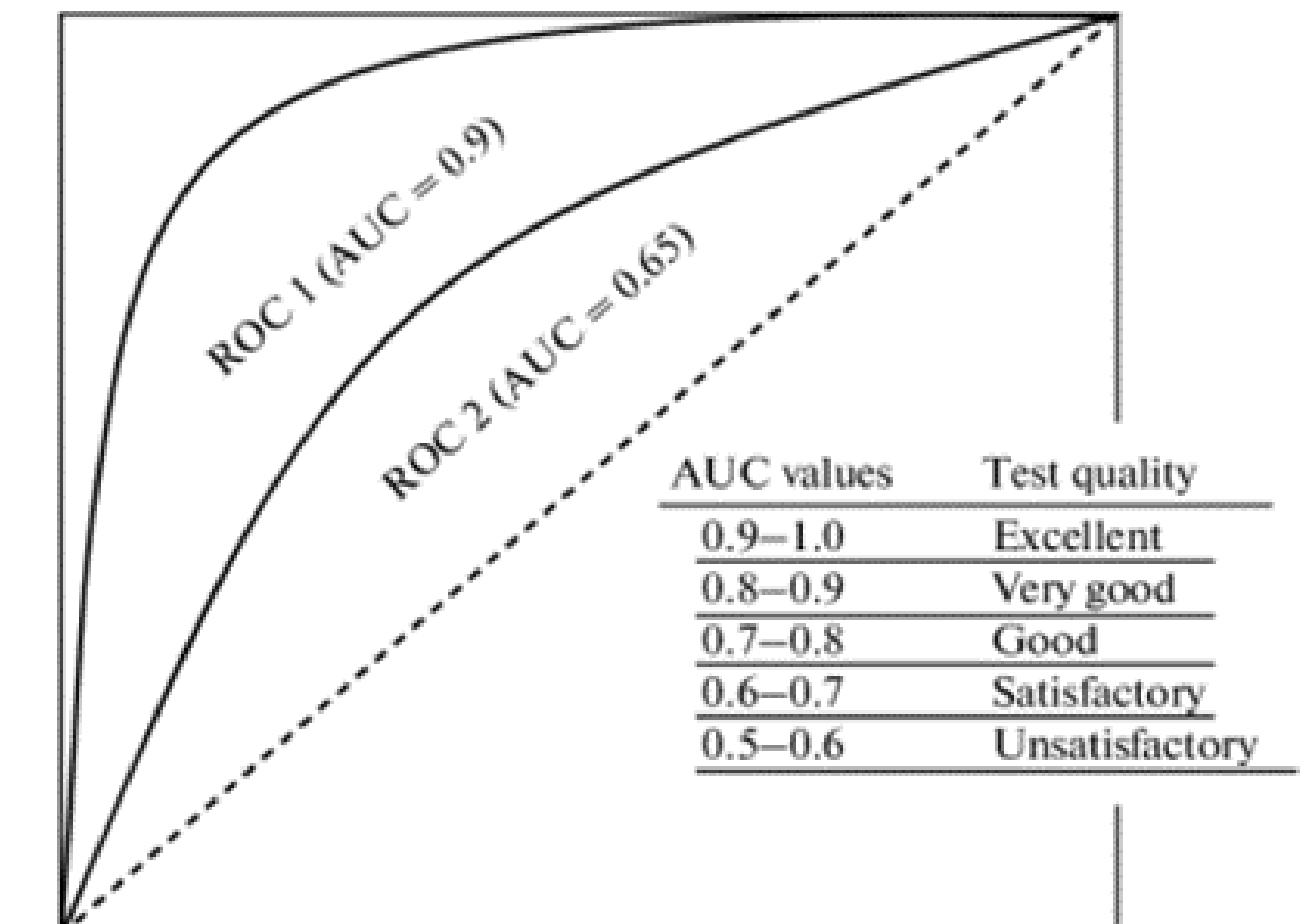


2.5.5 Các phương pháp đánh giá mô hình

2.5.5.3. AUC

Đường cong ROC biểu thị hiệu suất của mô hình phân loại trên toàn bộ phạm vi các ngưỡng dự đoán khác nhau. AUC là diện tích nằm dưới đường cong ROC, và giá trị AUC càng gần 1 thì mô hình càng tốt trong việc phân loại các lớp.

- **AUC = 0.5**
- **AUC > 0.5 và gần 1**
- **AUC = 1:**





CHƯƠNG 3. PHƯƠNG PHÁP THU THẬP VÀ XỬ LÝ DỮ LIỆU

3.1. Thu thập dữ liệu

3.1.1. Nguồn dữ liệu thời tiết :

- Bộ dữ liệu Weather là bộ dữ liệu được thu thập với mục đích phân tích và dự báo tình hình thời tiết để đưa ra giải pháp hữu ích phòng ngừa thời tiết xấu có thể xảy ra và giúp cho việc ra quyết định dễ dàng hơn. Bộ dữ liệu Weather được sử dụng trong bài báo cáo lần này được thu thập từ
- Bộ dữ liệu gồm có: 10 cột và 181960 dòng



3.1. Thu thập dữ liệu

3.1.2. Các thông số thời tiết quan trọng

STT	Tên cột	Kiểu dữ liệu	Ý nghĩa
1	Province	Object	Tỉnh hoặc thành phố
2	Max	Int64	Nhiệt độ tối đa trong ngày (°C)
3	Min	Int64	Nhiệt độ thiểu trong ngày (°C)
4	Wind	Int64	Tốc độ gió (km/h)
5	Wind_d	Object	Hướng gió
6	Rain	Float64	Lượng mưa (mm)
7	Humidi	Int64	Độ ẩm (%)
8	Cloud	Int64	Đám mây (%)
9	Pressure	Int64	Áp suất (bar)
10	Date	Object	Bản ghi ngày (yyyy-mm-dd)



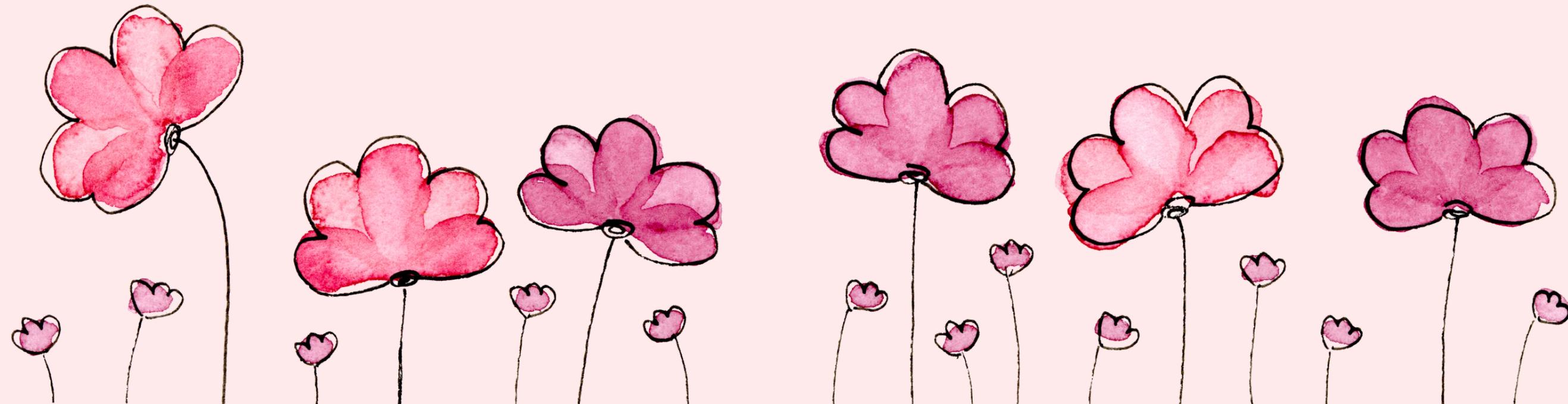
3.2. Xử lý dữ liệu

Xử lí dữ liệu là một phần quan trọng trong việc áp dụng mô hình học máy trong phân tích và dự đoán dữ liệu thời tiết. Quá trình này đảm bảo rằng dữ liệu đầu vào được tiền xử lí và biến đổi theo cách thích hợp để mô hình có thể hiểu và học từ dữ liệu một cách hiệu quả. Dưới đây là một số bước quan trọng trong việc xử lí dữ liệu cho đề tài "Áp dụng mô hình học máy trong phân tích và dự đoán dữ liệu thời tiết":

- 1.Thu thập dữ liệu
- 2.Tiền xử lí dữ liệu
- 3.Tạo các đặc trưng (features)
- 4.Chia dữ liệu thành tập huấn luyện và tập kiểm tra
- 5.Xây dựng mô hình học máy
- 6.Đánh giá và tinh chỉnh mô hình
- 7.Dự đoán và áp dụng



CHƯƠNG 4 : HỆ THỐNG PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO THỜI TIẾT



4.1.Yêu cầu đặt ra

Đề tài "Áp dụng mô hình học máy trong phân tích và dự đoán dữ liệu thời tiết" mang đến một số yêu cầu đặt ra quan trọng và đầy thách thức. Việc áp dụng mô hình học máy vào lĩnh vực dự đoán thời tiết có mục tiêu cung cấp các dự đoán chính xác và đáng tin cậy về tình hình thời tiết trong tương lai

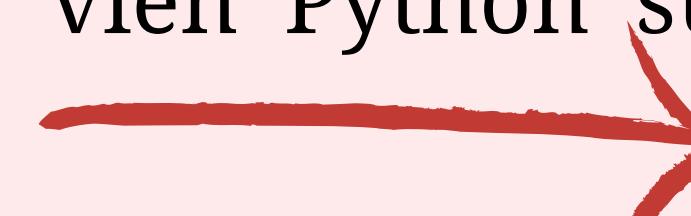


4.2.Tải và thao tác dữ liệu

- Dữ liệu từ Kaggle tiến hành đưa dữ liệu lên Jupyter note book để tiến hành phân tích dữ liệu



- Khai báo các thư viện Python sử dụng



```
1. Import Libraries [REDACTED]
```

```
import os
import time
import librosa
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from tqdm.notebook import trange,tqdm

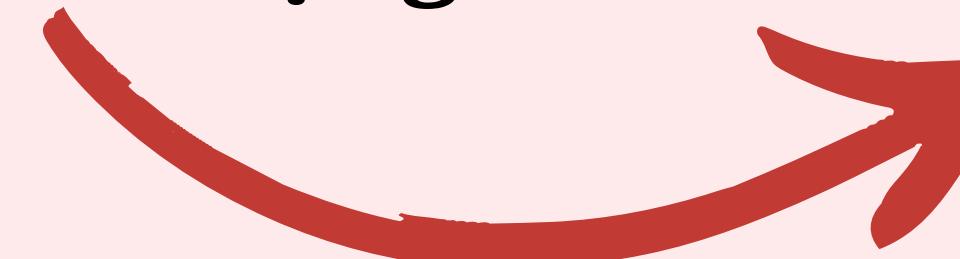
from sklearn.preprocessing import LabelEncoder, MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import *
from sklearnex import patch_sklearn, config_context

from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import ExtraTreesClassifier
from xgboost import XGBClassifier

from sklearn.model_selection import RepeatedStratifiedKFold, RandomizedSearchCV

import warnings
```

- Sau đó, thực hiện tải dữ liệu lên notebook để sử dụng



2.1 Reading dataset 📈

```
In [3]: df = pd.read_csv('weather.csv')
```

```
In [4]: df.head(10)
```

Out[4]:

	province	max	min	wind	wind_d	rain	humidi	cloud	pressure	date
0	Bac Lieu	27	22	17	NNE	6.9	90	71	1010	2009-01-01
1	Bac Lieu	31	25	20	ENE	0.0	64	24	1010	2010-01-01
2	Bac Lieu	29	24	14	E	0.0	75	45	1008	2011-01-01
3	Bac Lieu	30	24	30	E	0.0	79	52	1012	2012-01-01
4	Bac Lieu	31	25	20	ENE	0.0	70	24	1010	2013-01-01
5	Bac Lieu	28	23	14	ENE	0.0	75	55	1012	2014-01-01
6	Bac Lieu	29	23	10	ENE	0.4	75	42	1012	2015-01-01
7	Bac Lieu	32	24	22	ENE	0.0	63	9	1015	2016-01-01
8	Bac Lieu	30	24	20	ENE	0.5	76	35	1011	2017-01-01
9	Bac Lieu	29	23	16	E	0.0	70	33	1010	2018-01-01

```
print(f"Summary Of The Dataset with numerical columns :")  
df.describe()
```

Summary Of The Dataset with numerical columns :

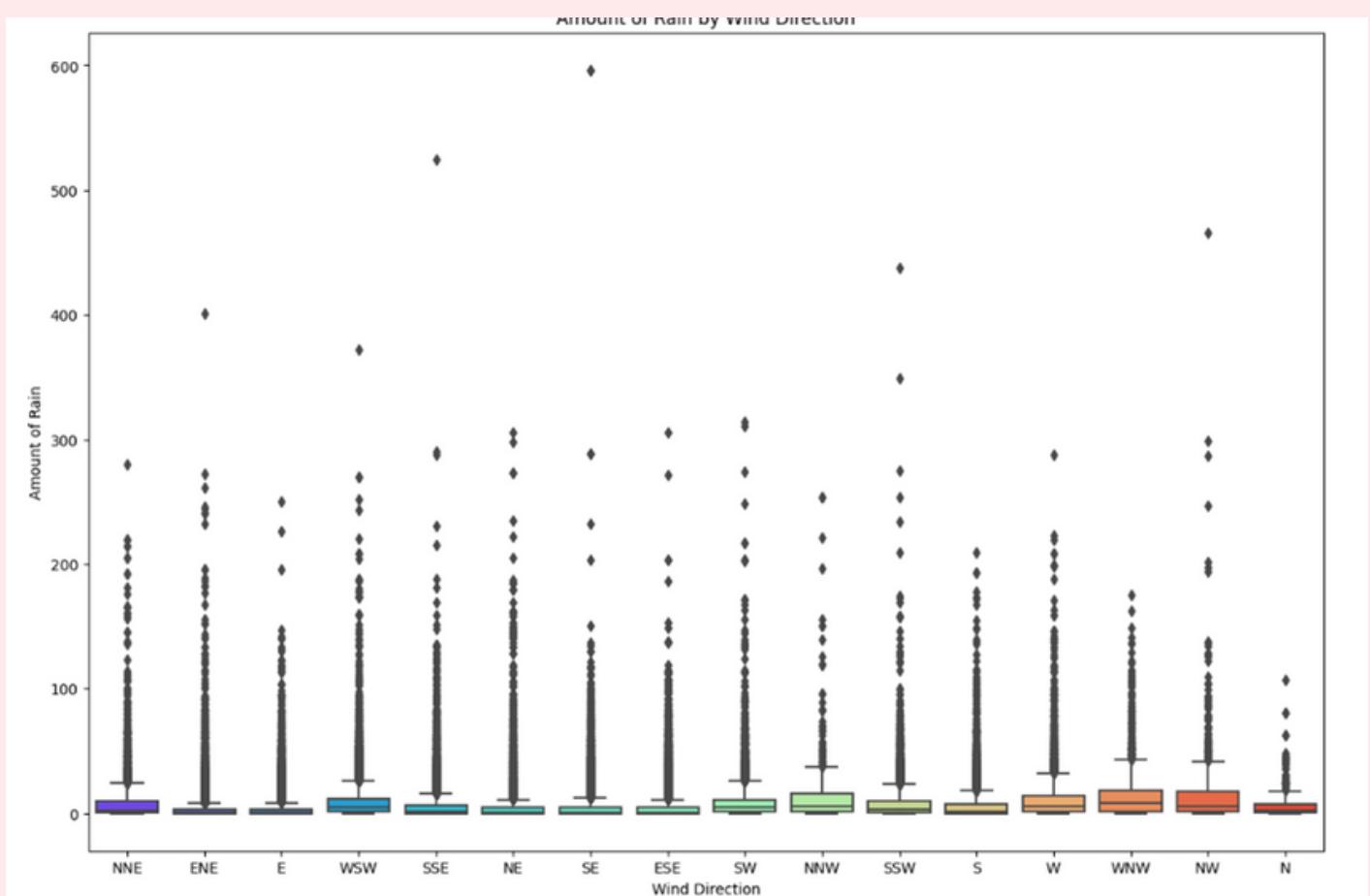
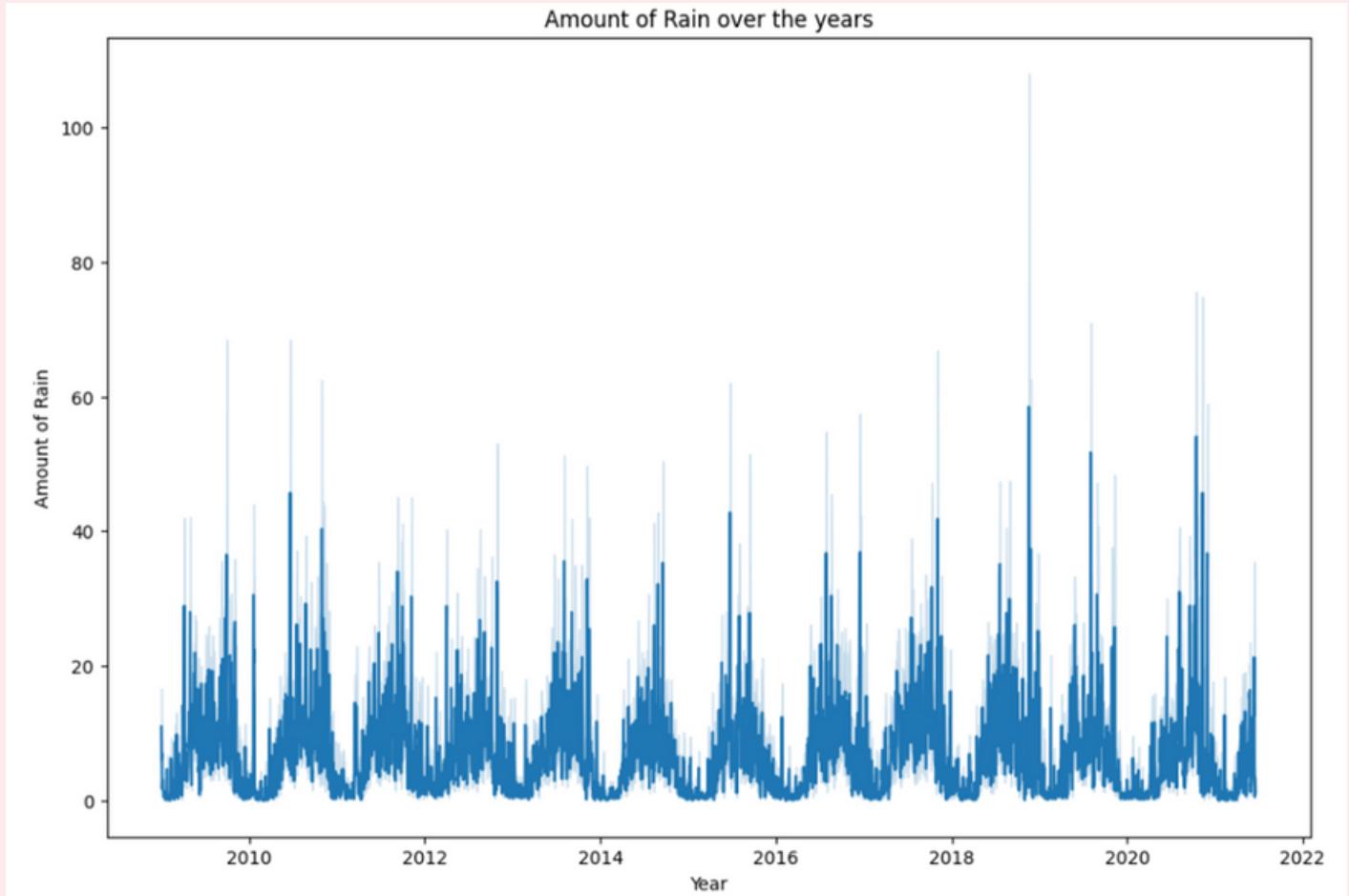
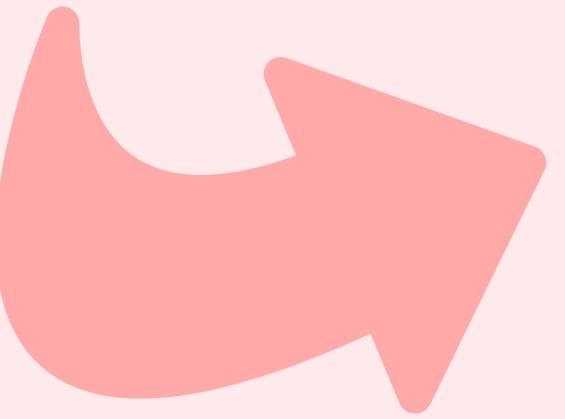
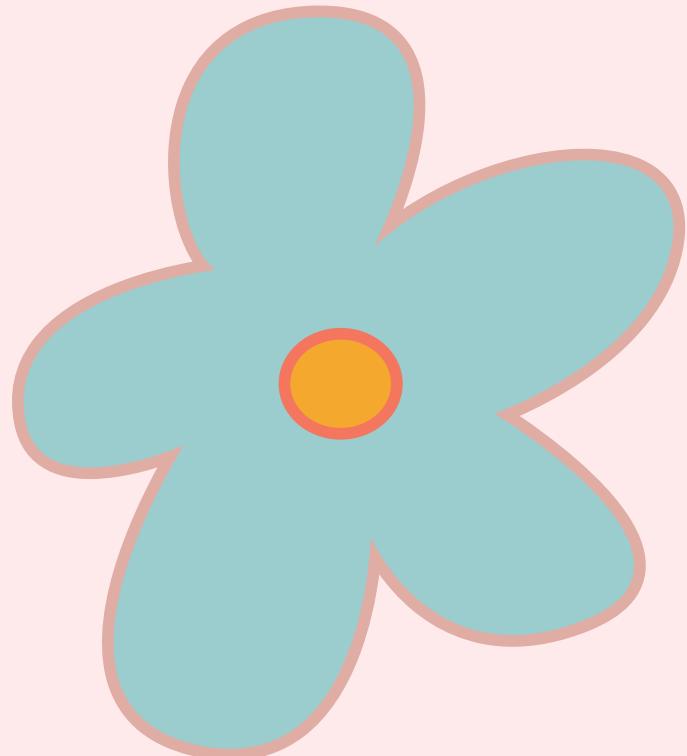
	max	min	wind	rain	humidi	cloud	pressure
count	181960.00	181960.00	181960.00	181960.00	181960.00	181960.00	181960.00
mean	29.84	23.28	11.04	6.57	77.08	41.72	1010.23
std	4.57	3.95	5.31	13.60	9.29	23.88	4.64
min	4.00	2.00	1.00	0.00	23.00	0.00	988.00
25%	28.00	21.00	7.00	0.10	71.00	23.00	1008.00
50%	31.00	24.00	10.00	1.80	78.00	38.00	1010.00
75%	33.00	26.00	14.00	7.50	83.00	58.00	1012.00
max	46.00	32.00	54.00	596.40	100.00	100.00	1038.00

- Thống kê các thông số của dữ liệu

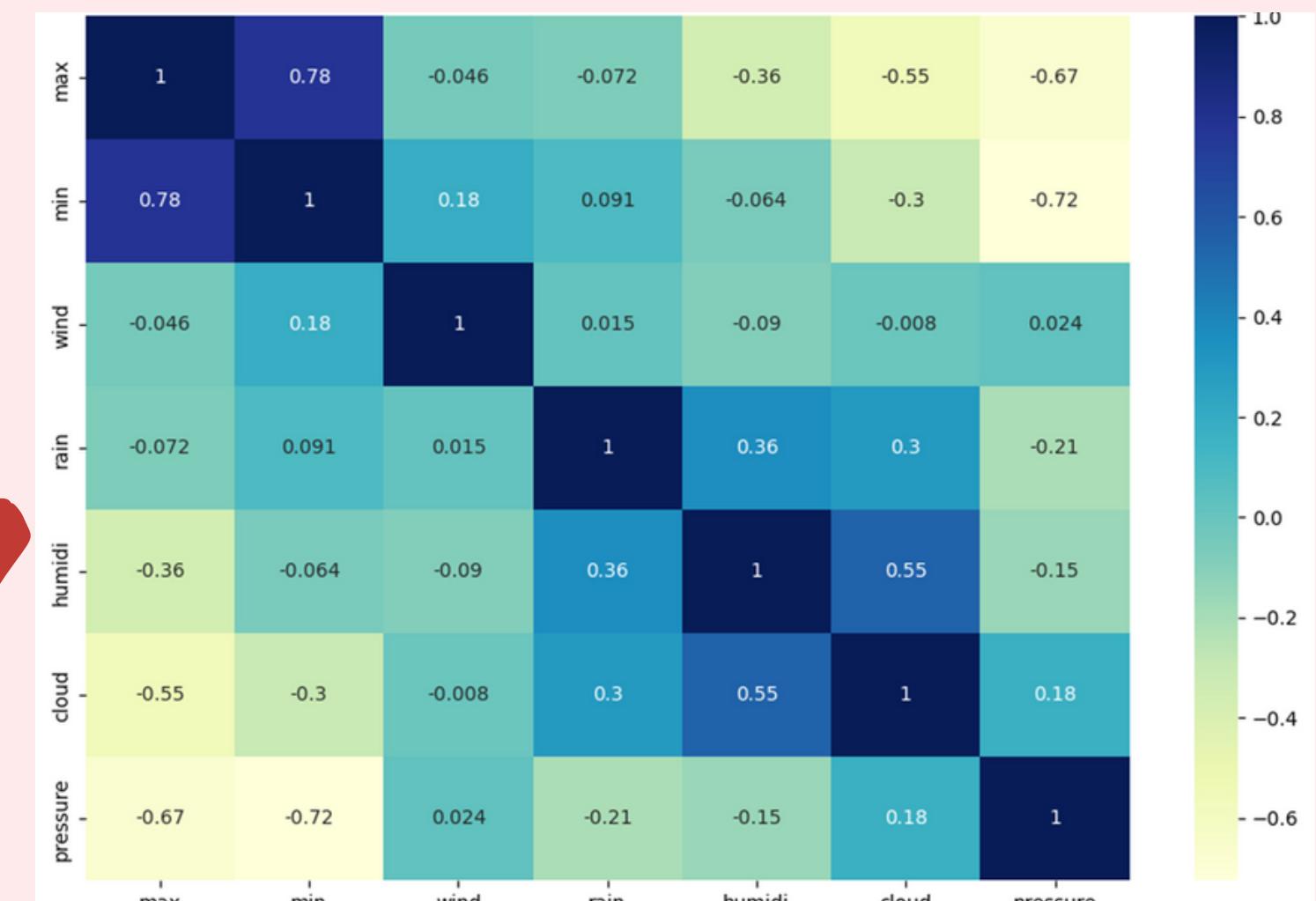
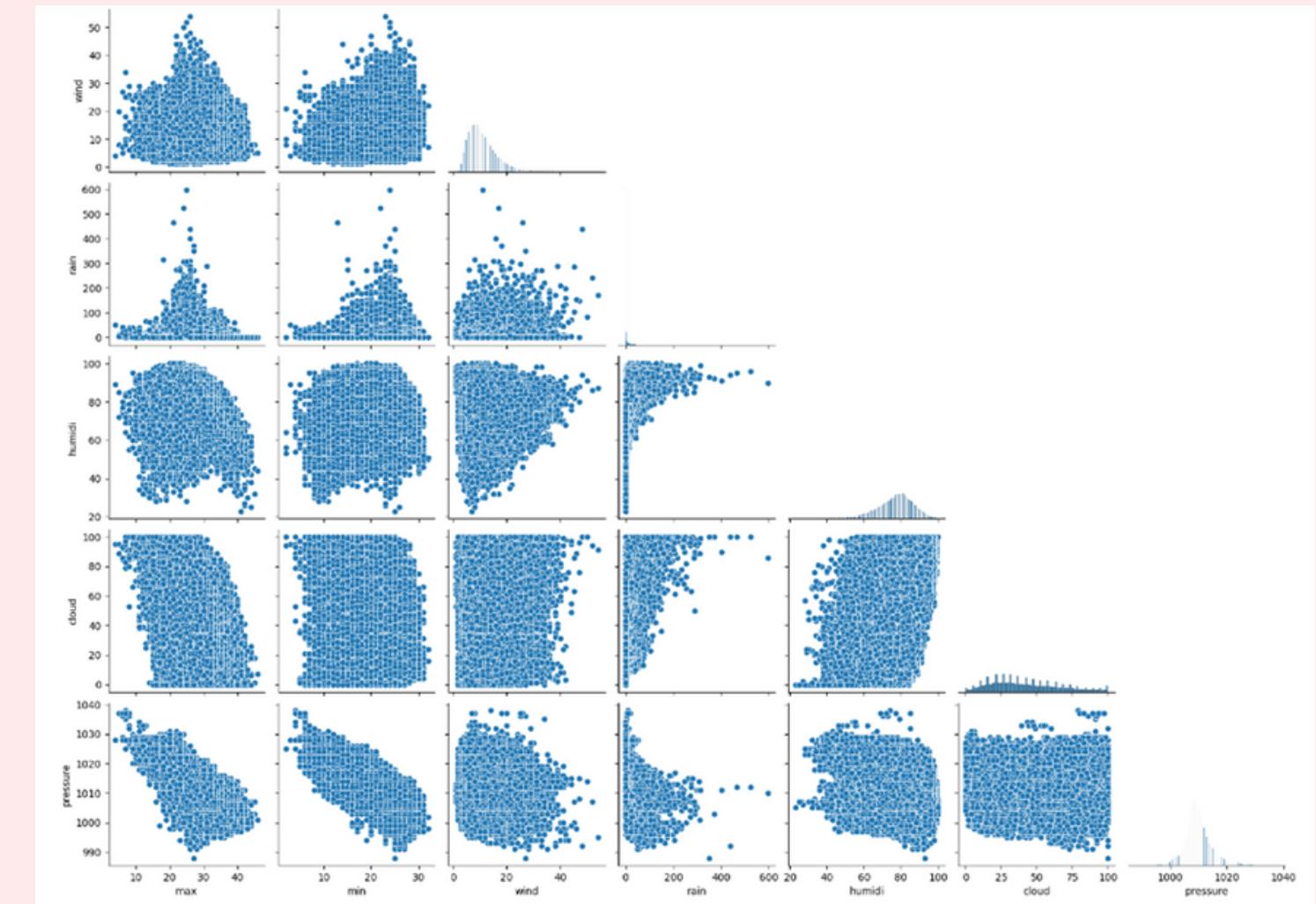


4.3. Xây dựng mô hình

Để xây dựng mô hình trên Jupyter notebook bằng ngôn ngữ python, đầu tiên cần nhập thư viện và các hàm cần thiết cho các mô hình xây dựng gồm K-Neighbors Classifier; Decision Tree Classifier và Random Forest.



Sau khi kiểm tra và phân tích với bộ dữ liệu. Em quyết định chọn hướng xây dựng mô hình học máy dựa trên các thuật toán Classification như: Logistic Regression, K-nearest Neighbors, Extra Trees, Naive Bayes classification, Decision Tree, Random Forest, Multilayer Perceptron, XGBoost.



	province	max	min	wind	wind_d	humidi	cloud	pressure
date								
2009-01-01	0	0.55	0.67	0.30	5	0.87	0.71	0.44
2010-01-01	0	0.64	0.77	0.36	1	0.53	0.24	0.44
2011-01-01	0	0.60	0.73	0.25	0	0.68	0.45	0.40
2012-01-01	0	0.62	0.73	0.55	0	0.73	0.52	0.48
2013-01-01	0	0.64	0.77	0.36	1	0.61	0.24	0.44
...
2016-12-28	27	0.57	0.70	0.13	11	0.68	0.50	0.46
2017-12-28	27	0.62	0.73	0.38	1	0.75	0.50	0.46
2018-12-28	27	0.52	0.73	0.15	1	0.88	0.75	0.42
2019-12-28	27	0.62	0.70	0.19	0	0.66	0.06	0.48
2020-12-28	27	0.60	0.73	0.15	2	0.79	0.43	0.42

Chạy vòng lặp for

```
for index in range(len(y)):
    if y[index] != '0.0':
        y[index] = 'Rain'
    else:
        y[index] = 'No Rain'
```

Tập huấn luyện train và tập kiểm tra test

Chuyển đổi từ object sang numeric

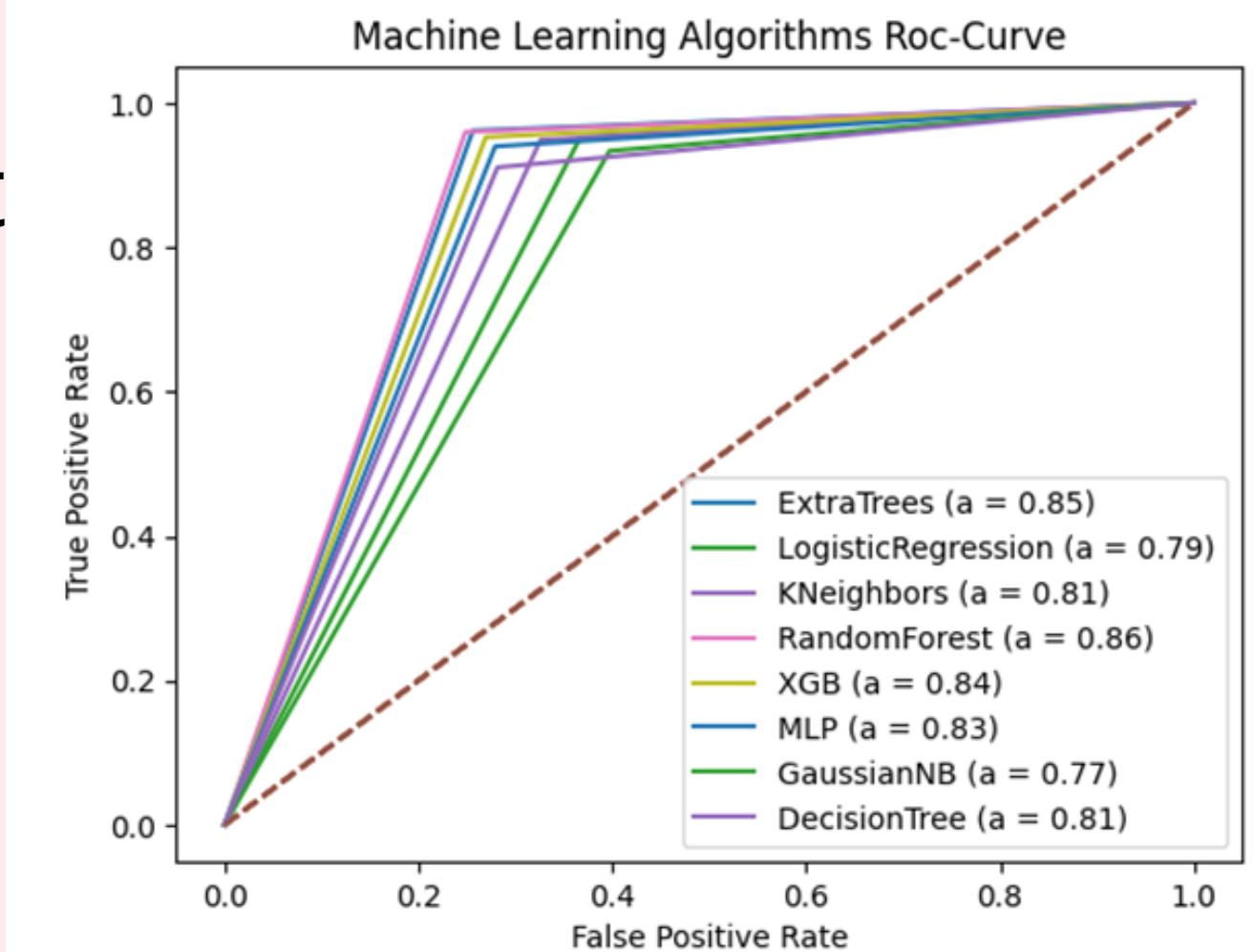
	Model	Accuracy(Train)	Accuracy(Test)	F1(Train)	F1(Test)	Precision(Train)	Precision(Test)	Recall(Train)	Recall(Test)	Log_loss(Train)
	GaussianNB	0.856499	0.859914	0.856499	0.859914	0.856499	0.859914	0.856499	0.859914	0.345225
	DecisionTreeClassifier	1.000000	0.867956	1.000000	0.867956	1.000000	0.867956	1.000000	0.867956	0.000000
	LogisticRegression	0.879236	0.881348	0.879236	0.881348	0.879236	0.881348	0.879236	0.881348	0.290812
	KNeighborsClassifier	0.917360	0.887576	0.917360	0.887576	0.917360	0.887576	0.917360	0.887576	0.171333
	MLPClassifier	0.890651	0.891038	0.890651	0.891038	0.890651	0.891038	0.890651	0.891038	0.258806
	XGBClassifier	0.914432	0.903202	0.914432	0.903202	0.914432	0.903202	0.914432	0.903202	0.203193
	ExtraTreesClassifier	1.000000	0.913223	1.000000	0.913223	1.000000	0.913223	1.000000	0.913223	0.000000
	RandomForestClassifier	0.999992	0.913369	0.999992	0.913369	0.999992	0.913369	0.999992	0.913369	0.057682

```
Accuracy_set.sort_values(by='Accuracy(Test)').style.background_gradient(cmap= plt.cm.Blues)
```

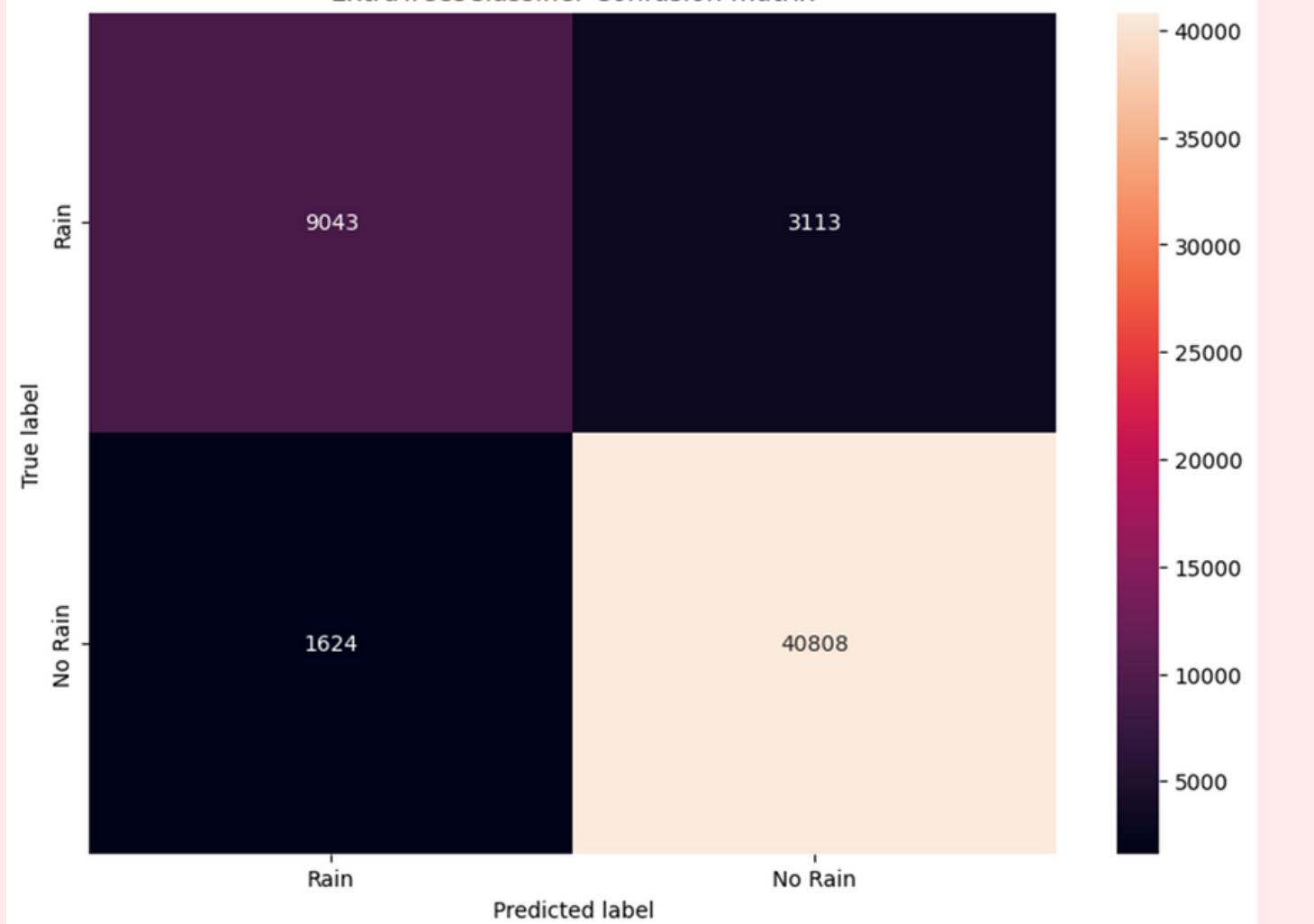
Model	Accuracy(Train)	Accuracy(Test)	F1(Train)	F1(Test)	Precision(Train)	Precision(Test)	Recall(Train)	Recall(Test)	Log_loss(Train)
GaussianNB	0.856499	0.859914	0.856499	0.859914	0.856499	0.859914	0.856499	0.859914	0.345225
DecisionTreeClassifier	1.000000	0.867956	1.000000	0.867956	1.000000	0.867956	1.000000	0.867956	0.000000
LogisticRegression	0.879236	0.881348	0.879236	0.881348	0.879236	0.881348	0.879236	0.881348	0.290812
KNeighborsClassifier	0.917360	0.887576	0.917360	0.887576	0.917360	0.887576	0.917360	0.887576	0.171333
MLPClassifier	0.890651	0.891038	0.890651	0.891038	0.890651	0.891038	0.890651	0.891038	0.258806
XGBClassifier	0.914432	0.903202	0.914432	0.903202	0.914432	0.903202	0.914432	0.903202	0.203193
ExtraTreesClassifier	1.000000	0.913223	1.000000	0.913223	1.000000	0.913223	1.000000	0.913223	0.000000
RandomForestClassifier	0.999992	0.913369	0.999992	0.913369	0.999992	0.913369	0.999992	0.913369	0.057682

Kết quả của quá trình chạy train và test

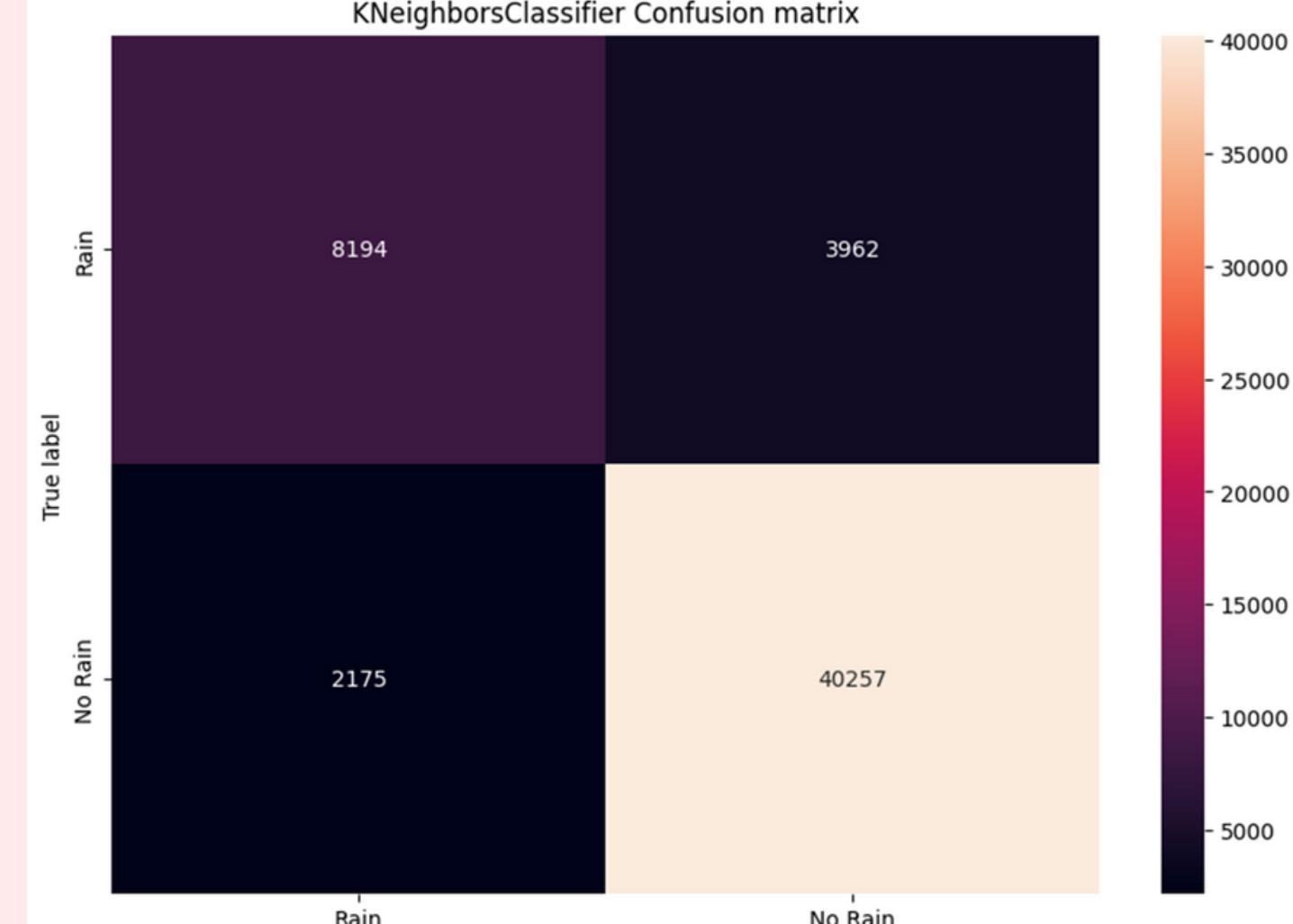
Kết quả của đồ thị ROC →



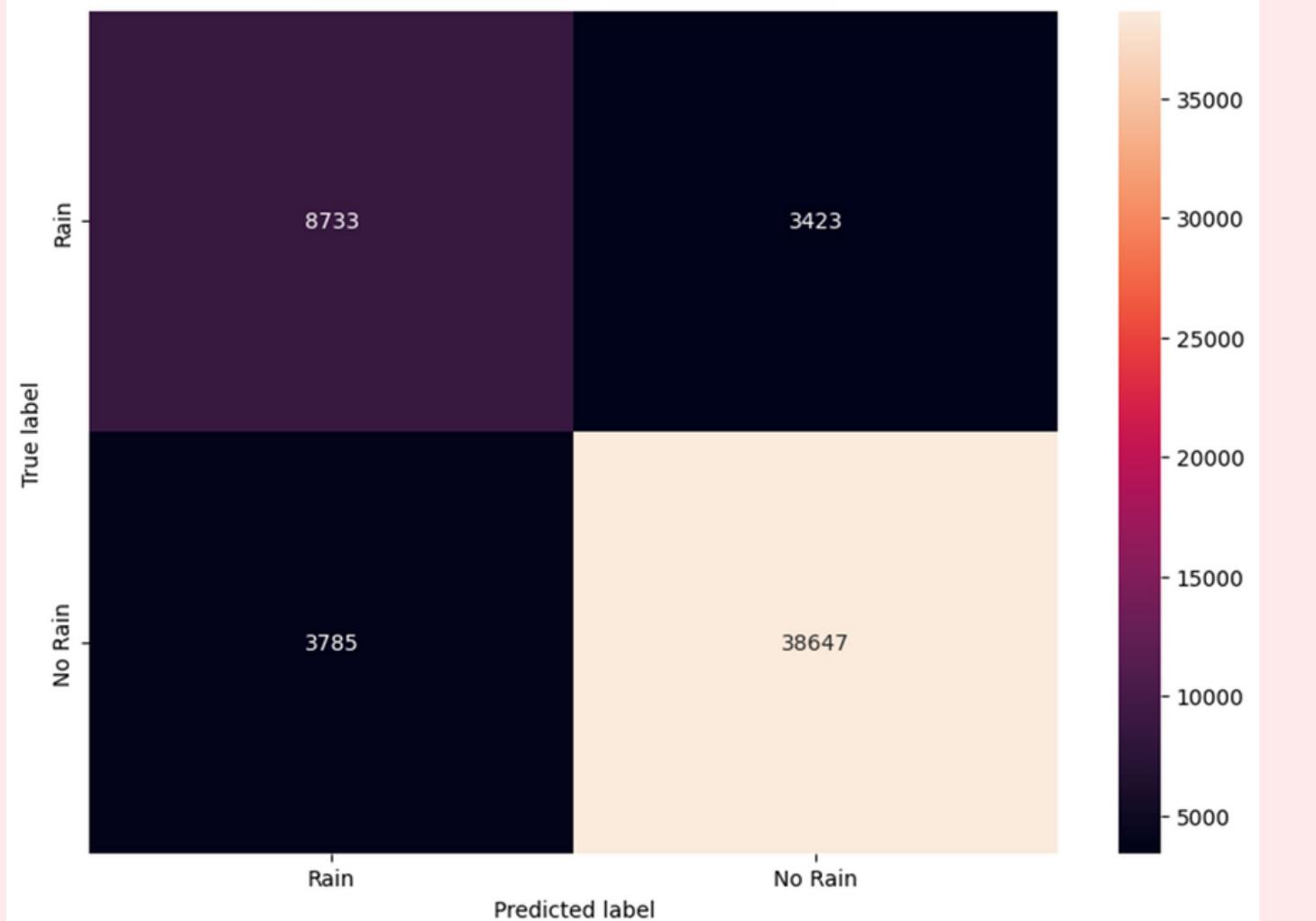
ExtraTreesClassifier Confusion matrix



KNeighborsClassifier Confusion matrix

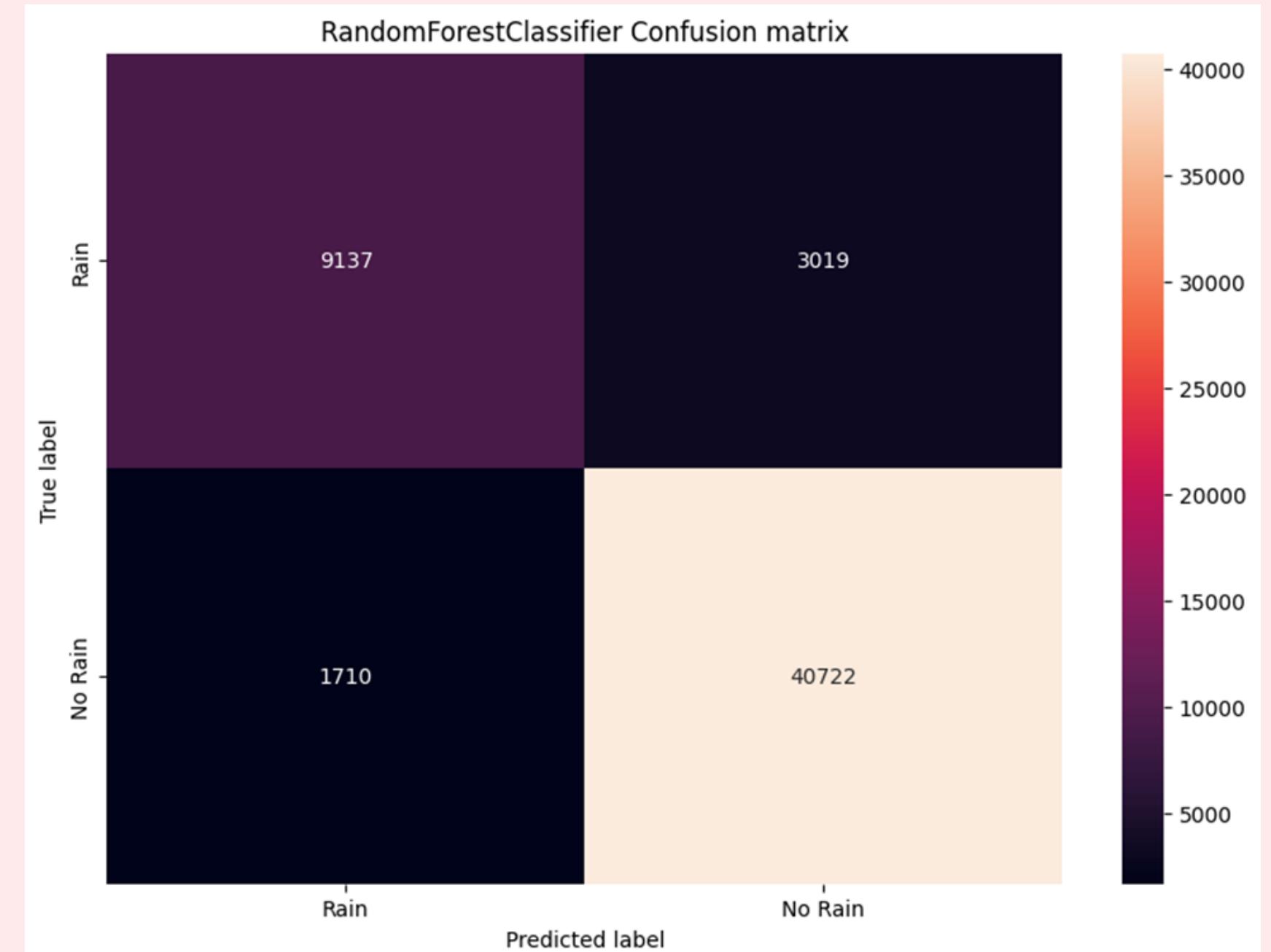
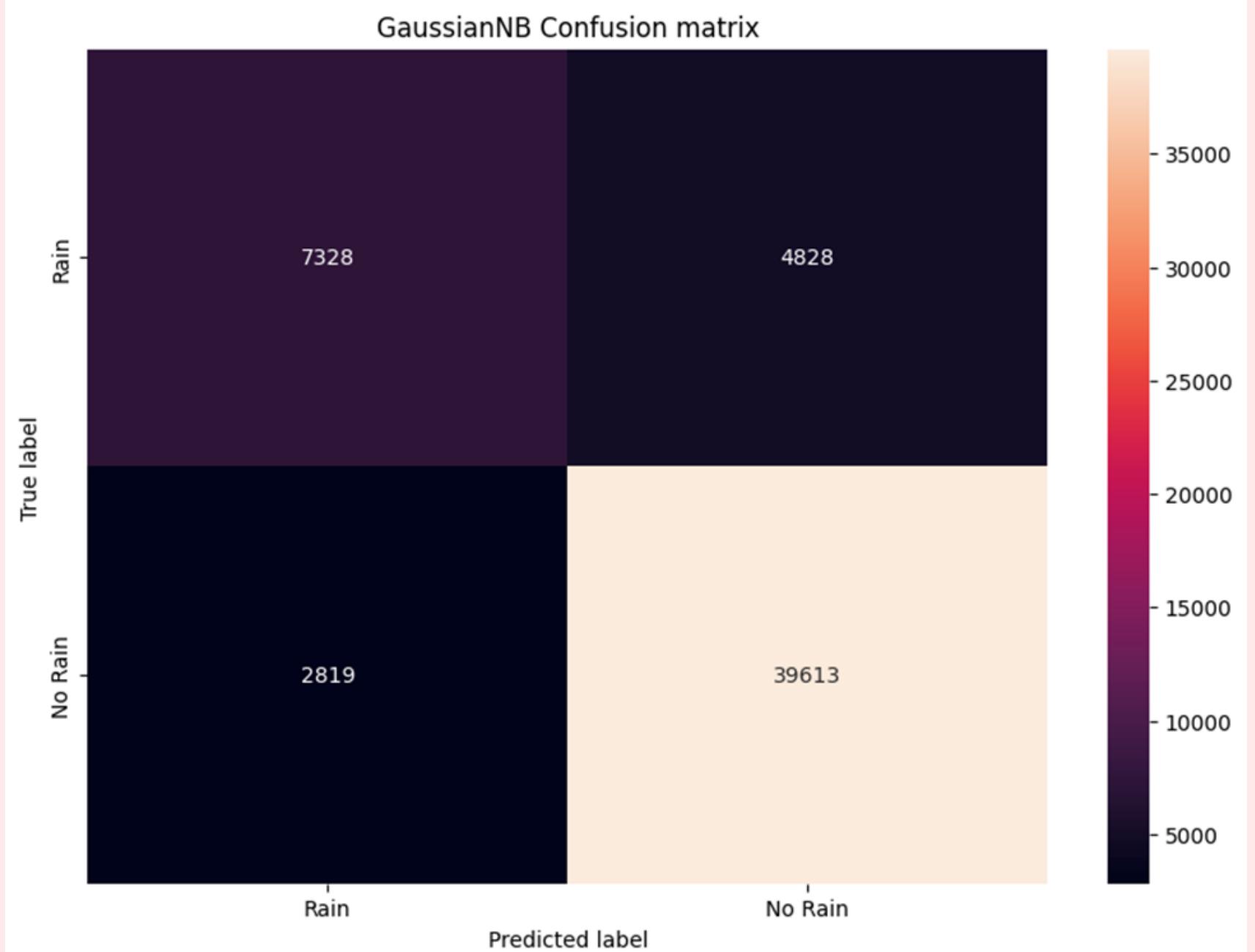


DecisionTreeClassifier Confusion matrix

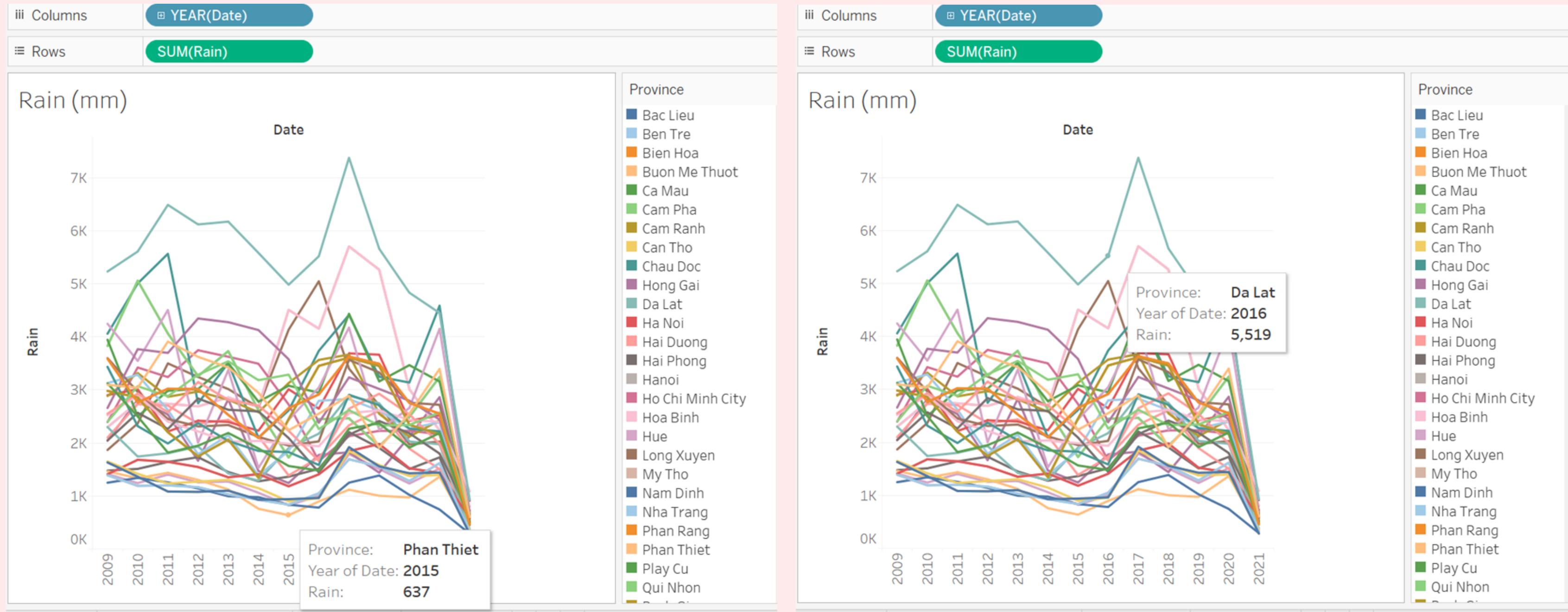


MLPClassifier Confusion matrix





4.4.Trực quan hóa và phân tích tình hình thời tiết



Biểu đồ tổng lượng mưa ở các Tỉnh trong Nước Việt Nam

Filters

YEAR(Date): 2021

Marks

<input type="checkbox"/> Automatic
Color
Size
Label
Detail
Tooltip
AVG(Max)
AVG(Max)
Province
AVG(Max)

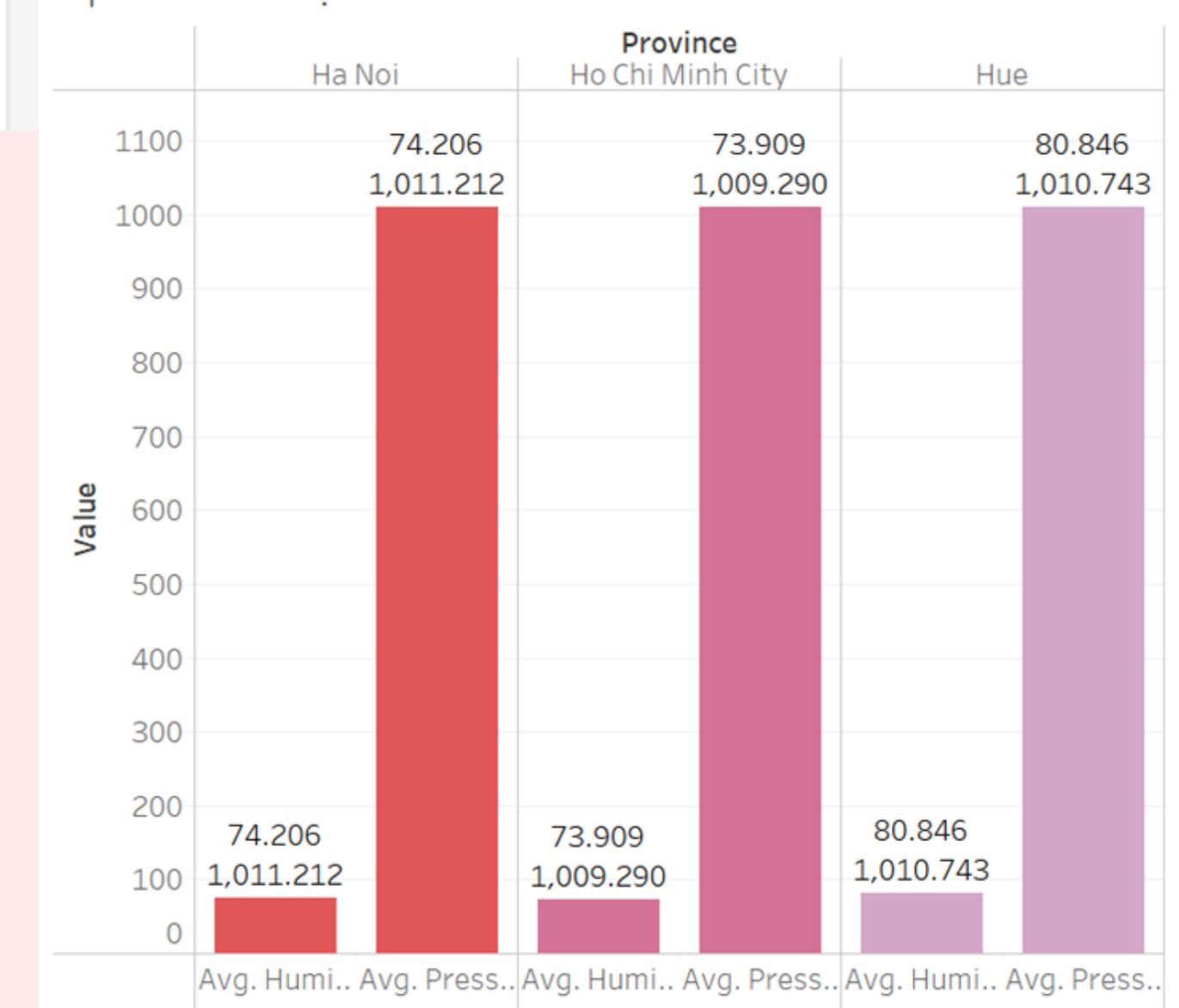
Nhiệt độ Max(°C)

Bien Hoa 33.787	Long Xuyen 32.781	Soc Trang 31.627	Vung Tau 28.272	Qui Nhon 27.817	Tuy Hoa 27.817	Tam Ky 27.763	Ha Noi 27.692
Ho Chi Minh City 32.615	Vinh Long 32.615	Buon Me Thuot	Hanoi 27.692	Phan Rang	Hoa Binh		Nam Dinh
Chau Doc 33.183	Can Tho 32.491	Rach Gia 30.592	Nha Trang 27.645				
Tan An 33.118	Ca Mau 32.112	Play Cu 30.260	Viet Tri 27.497	Hai Duong 26.746	Hai Phong 26.314	Da Lat 26.231	
Ben Tre 32.994	Tra Vinh 31.686	Phan Thiet 30.041	Yen Bai 27.497	Vinh 26.367	Uong Bi 25.432	Cam Pha	
My Tho 32.994	Bac Lieu 31.627	Hue 28.379	Cam Ranh 27.260	Thai Nguyen	Hong Gai 24.071		

AVG(Max)

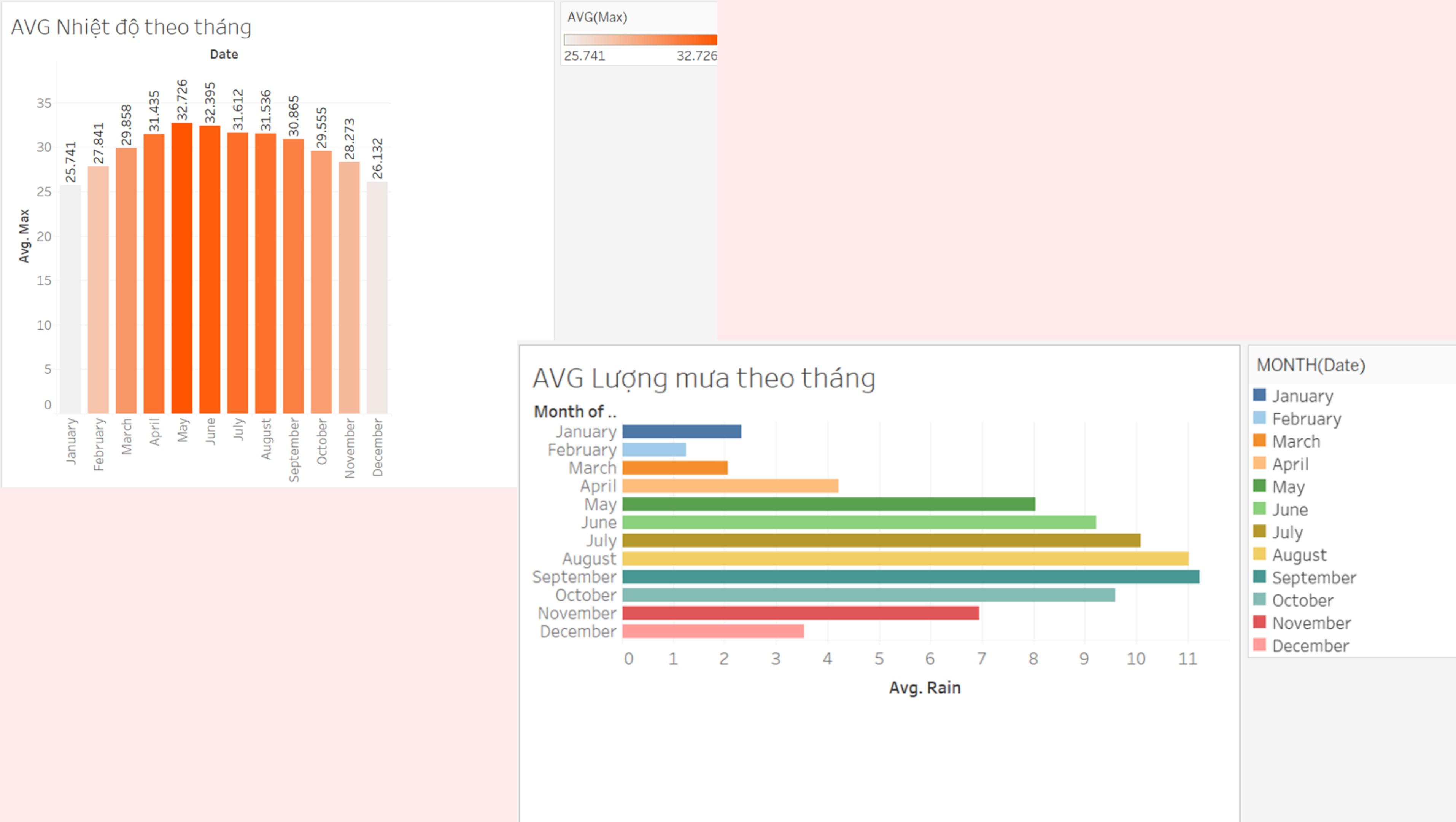
23.988 33.787

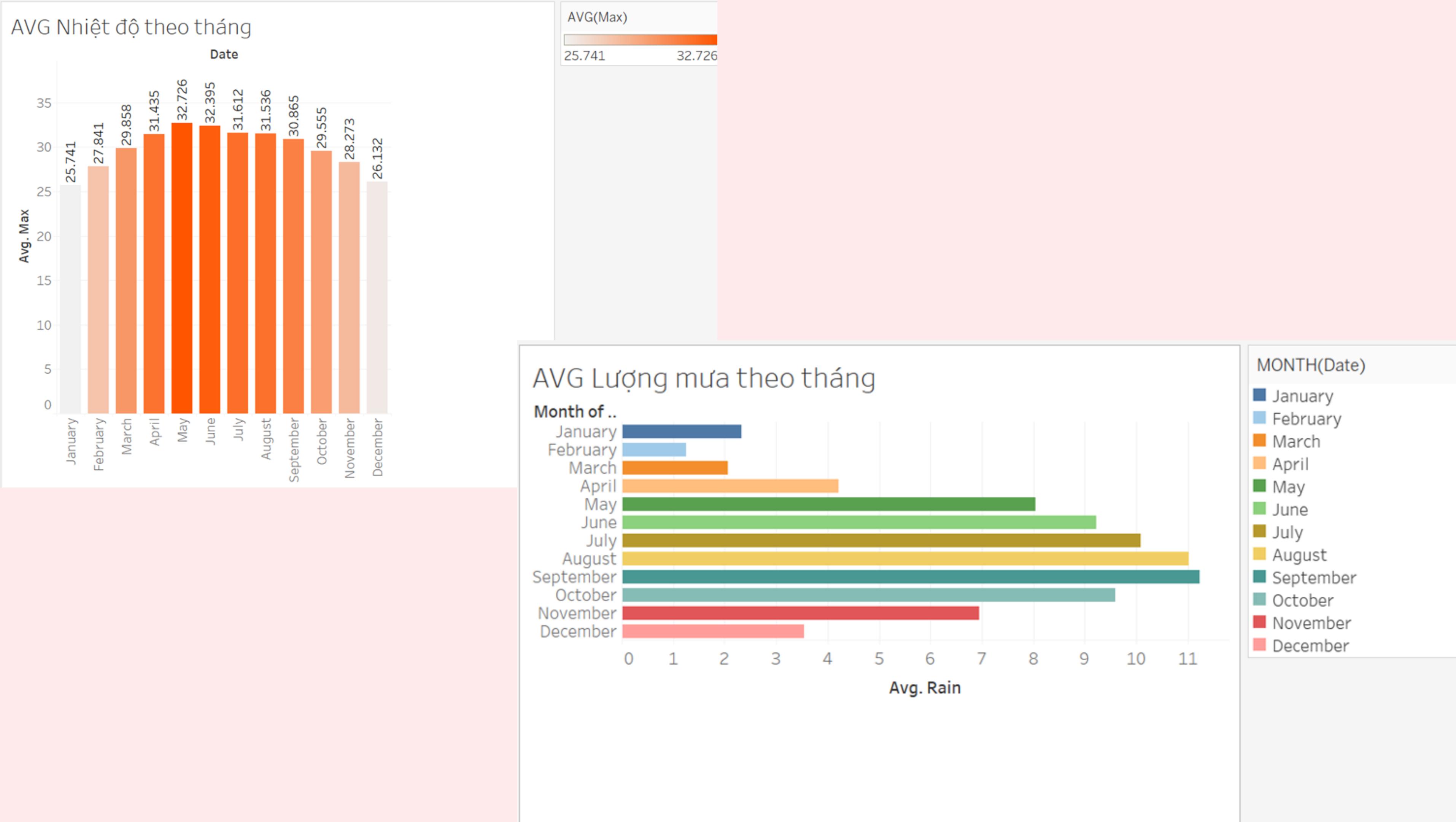
Áp suất & Độ ẩm



Province

- Ha Noi
- Ho Chi Minh City
- Hue





4.5. Đề xuất và ứng dụng

4.5.1.Khuyến khích sử dụng thông tin thời tiết đời sống



4.5. Đề xuất và ứng dụng

4.5.2. Ứng dụng thông tin thời tiết vào các quyết định kinh doanh

- Quản lý nguồn lực hiệu quả:
- Chiến lược tiếp thị đích đáng:
- Quản lý chuỗi cung ứng:
- Nâng cao trải nghiệm khách hàng:
- Quản lý rủi ro và bảo hiểm



KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

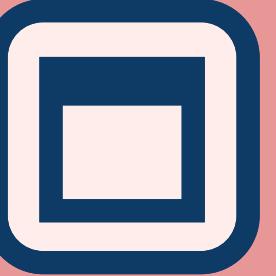
- **Đạt được:**
- **Về đề tài:**

Cải thiện dự báo thời tiết: Sử dụng mô hình học máy đã giúp cải thiện độ chính xác của dự báo thời tiết. Nhờ khả năng xử lý dữ liệu phức tạp và nhận dạng mẫu ẩn, mô hình học máy đã đóng góp vào việc tạo ra những dự báo thời tiết có tính chính xác cao hơn và phản ánh tốt hơn sự biến đổi thời tiết.

Nghiên cứu thành công và trình bày khái quát các nội dung về Data Analyst và nắm rõ các công việc liên quan tới vị trí DA, các công cụ xử lí và phân tích dữ liệu như Tableau Prep Builder, Tableau và ứng dụng Data Mining vào xây dựng các mô hình dữ liệu

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

- **Hạn chế:**
 - Bộ dữ liệu mẫu không có các trường thời gian nên việc phân tích chưa đạt được những mong muốn liên quan tới tham số thời gian, vì thế việc đánh giá xu hướng, so sánh dữ liệu chưa thật sự hiệu quả
- **Hướng phát triển**
 - Trau dồi thêm các kiến thức chuyên ngành liên quan khi thực hiện các công việc liên quan tới dữ liệu tương ứng.
 - Phát triển công việc trong lĩnh vực dữ liệu



Thanks Y'all

