TRƯỜNG ĐẠI HỌC KINH TẾ

**KHOA THỐNG KÊ – TIN HỌC**



**BÁO CÁO THỰC TẬP NGHỀ NGHIỆP**

**NGÀNH HỆ THỐNG THÔNG TIN QUẢN LÝ**

**CHUYÊN NGÀNH QUẢN TRỊ HỆ THỐNG THÔNG TIN**

**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**

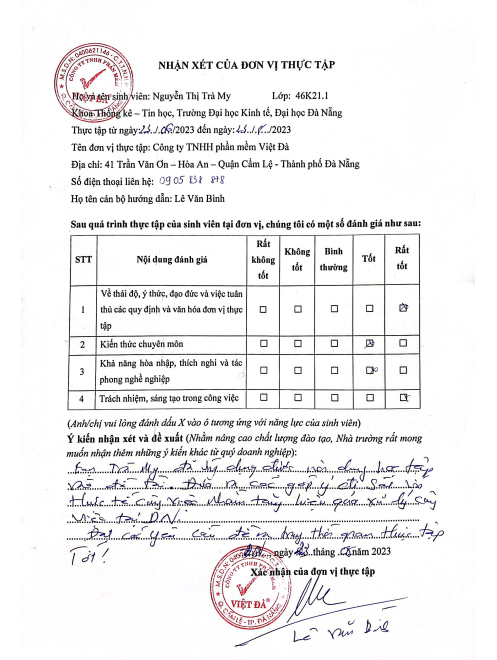
Đơn vị thực tập : **Công Ty TNHH phần mềm Việt Đà**

Cán bộ hướng dẫn : **Lê Văn Bình**

Giảng viên hướng dẫn : **TS. Phan Đình Vấn**

**Đà Nẵng, 8/202****3**

**NHẬN XÉT CỦA ĐƠN VỊ THỰC TẬP**

****

# LỜI CẢM ƠN

Lời đầu tiên, em xin gửi lời cảm ơn chân thành và sự biết ơn đến tất cả những người đã hỗ trợ em về kiến thức, chuyên môn, tinh thần và tạo điều kiện cho em hoàn thiện tốt đề tài báo cáo **“ Áp dụng mô hình máy học trong phân tích và và dự đoán dữ liệu thời tiết”.**

Em xin chân thành cảm ơn Thầy TS. Phan Đình Vấn đã trực tiếp hướng dẫn em thực hiện đề tài báo cáo thực tập nghề nghiệp. Thầy đã tận tình chỉ dẫn, nhận xét, giúp đỡ và hỗ trợ em trong quá trình thực hiện và hoàn thành đề tài báo cáo.

Em cũng xin gửi đến quý thầy, cô giáo trong khoa Thống Kê – Tin Học trường Đại Học Kinh Tế – Đại học Đà Nẵng lời cảm ơn chân thành vì đã tận tâm giảng dạy và truyền đạt những kiến thức, kinh nghiệm quý báu cho em.

Cuối cùng em xin cảm ơn Anh Lê Văn Bình và các anh chị Công Ty TNHH phần mềm Việt Đà.Trong quá trình thực tập các anh chị đã tận tình giúp đỡ về chuyên môn, nghiệp vụ, cung cấp những tài liệu, định hướng và tạo mọi điều kiện thuận lợi để giúp em có thể hoàn thành bài báo cáo thực tập.

Vì thời gian và kiến thức còn hạn hẹp nên bài báo cáo không thể tránh khỏi những thiếu sót, rất mong sự góp ý của công ty, quý thầy cô và các bạn, để em rút kinh nghiệm và hoàn thành tốt hơn.

Em xin chân thành cảm ơn!

# LỜI CAM ĐOAN

Em xin cam đoan đề tài **“ Áp dụng mô hình máy học trong phân tích và và dự đoán dữ liệu thời tiết”** là kết quả từ quá trình thực tập tại doanh nghiệp của em dưới sự hướng dẫn của giảng viên hướng dẫn: TS. Phan Đình Vấn và cán bộ hướng dẫn là anh Lê Văn Bình. Ngoài ra không có bất cứ sự sao chép nào của người khác.

Đề tài, nội dung báo cáo là sản phẩm mà em đã được thực hành trong quá trình thực tập tại doanh nghiệp. Các dữ liệu, kết quả trình bày trong báo cáo là hoàn toàn trung thực, em xin chịu hoàn toàn trách nhiệm với bộ môn và nhà trường nếu có bất kỳ vấn đề nào xảy ra.

# MỤC LỤC

[LỜI CẢM ƠN ii](#_Toc143672209)

[LỜI CAM ĐOAN iii](#_Toc143672210)

[MỤC LỤC iv](#_Toc143672211)

[DANH MỤC HÌNH ẢNH vi](#_Toc143672212)

[DANH MỤC BẢNG BIỂU vii](#_Toc143672213)

[DANH MỤC CÁC TỪ VIẾT TẮT viii](#_Toc143672214)

[LỜI MỞ ĐẦU 1](#_Toc143672215)

[CHƯƠNG 1. TỔNG QUAN VỀ ĐỀ TÀI VÀ CÔNG TY TNHH PHẦN MỀM VIỆT ĐÀ 3](#_Toc143672216)

[1.1.Giới thiệu tổng quát về công ty TNHH phần mềm Việt Đà 3](#_Toc143672217)

[1.1.1.Tổng quan về doanh nghiệp 3](#_Toc143672218)

[1.1.2.Tầm nhìn, sứ mệnh và ngành nghề kinh doanh của công ty 5](#_Toc143672219)

[1.2.Tổng quan về đề tài 11](#_Toc143672220)

[1.2.1.Vấn đề thực tiễn 11](#_Toc143672221)

[CHƯƠNG 2.CƠ SỞ LÝ THUYẾT PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO THỜI TIẾT 13](#_Toc143672222)

[2.1.Giới thiệu về Data Analyst 13](#_Toc143672223)

[2.1.1.Khái niệm: 14](#_Toc143672224)

[2.1.2.Công việc của Data Analyst trong doanh nghiệp: 14](#_Toc143672225)

[2.2.Tổng quan về dữ liệu thời tiết và kinh doanh 16](#_Toc143672226)

[2.2.1 Khái niệm về dữ liệu thời tiết: 17](#_Toc143672227)

[2.2.2 Tầm quan trọng của dữ liệu thời tiết trong đời : 17](#_Toc143672229)

[2.2.3 Lợi ích và ứng dụng của phân tích dữ liệu thời tiết trong đời sống: 18](#_Toc143672230)

[2.3.1. Khái niệm Tableau Prep Builder: 19](#_Toc143672231)

[2.3.2. Đặc điểm nổi bật Tableau Prep Builder: 20](#_Toc143672232)

[2.4.1. Khái niệm Tableau: 21](#_Toc143672233)

[2.4.2. Các tính năng nổi bật của Tableau: 22](#_Toc143672234)

[2.4.3. Các loại biểu đồ thường dùng: 22](#_Toc143672235)

[2.5.1 Khái niệm: 23](#_Toc143672236)

[2.5.2 Vai trò của việc khai phá dữ liệu: 24](#_Toc143672237)

[2.5.3 Các kỹ thuật khai phá dữ liệu: 24](#_Toc143672238)

[2.5.4 Các mô hình khai phá dữ liệu: 25](#_Toc143672239)

[2.5.5 Các phương pháp đánh giá mô hình 28](#_Toc143672240)

[CHƯƠNG 3. PHƯƠNG PHÁP THU THẬP VÀ XỬ LÝ DỮ LIỆU 31](#_Toc143672241)

[3.1.Thu thập dữ liệu 32](#_Toc143672242)

[3.1.1. Nguồn dữ liệu thời tiết : 32](#_Toc143672243)

[3.1.2. Các thông số thời tiết quan trọng 32](#_Toc143672244)

[3.2.Xử lý dữ liệu 32](#_Toc143672245)

[CHƯƠNG 4 : HỆ THỐNG PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO THỜI TIẾT 34](#_Toc143672246)

[4.1.Yêu cầu đặt ra 34](#_Toc143672247)

[4.2.Tải và thao tác dữ liệu 35](#_Toc143672248)

[4.3. Xây dựng mô hình 37](#_Toc143672249)

[4.4.Trực quan hóa và phân tích tình hình thời tiết 49](#_Toc143672250)

[4.5. Đề xuất và ứng dụng 57](#_Toc143672251)

[4.5.1.Khuyến khích sử dụng thông tin thời tiết đời sống 57](#_Toc143672252)

[4.5.2. Ứng dụng thông tin thời tiết vào các quyết định kinh doanh 57](#_Toc143672253)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 58](#_Toc143672254)

[TÀI LIỆU THAM KHẢO 60](#_Toc143672255)

# DANH MỤC HÌNH ẢNH

*Hình 1.1 Sơ đồ cơ cấu tổ chức của Công Ty*

*Hình 2.1.2 Các công việc của data analyst*

Hình 2.3.1 Tableau Prep Builder

Hình 2.4.1 Tableau

Hình 2.5.5.3 Minh họa AUC

Hình 4.2 Khai báo các thư viên Python sử dụng

Hình 4.3 Dữ liệu sau khi tải lên notebook

*Hình 4.4 Xem dữ liệu có trong các côt*

*Hình 4.5 Thống kê các thông số của dữ liệu*

*Hình 4.6 Minh họa Correlation Matrix*

*Hình 4.7 Lượng mưa qua các năm*

*Hình 4.8 Lượng mưa theo hướng gió*

*Hình 4.9 Chuyển đổi từ object sáng numeric*

*Hình 4.10 chạy vòng lặp for*

*Hình 4.11 Tập huấn luyện train và tập kiểm tra test*

*Hình 4.12 Kết quả của quá trình chạy train và test*

*Hình 4.13 Kết quả của đồ thị ROC*

*Hình 4.14 Ma trận Confusion Matrix của mô hình ExtraTreesClassifier*

*Hình 4.15 Ma trận Confusion Matrix của mô hình KNeighbors Classifier*

*Hình 4.16 Ma trận Confusion Matrix của mô hình Decision TreeClassifier*

*Hình 4.17 Ma trận Confusion Matrix của mô hình MLP Classifier*

*Hình 4.18 Ma trận Confusion Matrix của mô hình GaussianNB Confusion*

*Hình 4.19 Ma trận Confusion Matrix của mô hình Random Forest 1Classifier*

*Hình 4.4 Biểu đồ tổng lượng mưa ở các Tỉnh trong Nước Việt Nam*

*Hình 4.5 Biểu đồ trung bình nhiệt độ cao nhất ở các tỉnh*

*Hình 4.6 Biểu đồ Áp suất và độ ẩm của các 3 Tỉnh*

*Hình 4.7 Biểu đồ nhiệt độ max theo tháng*

*Hình 4.8 Biểu đồ Lượng mưa theo tháng*

# DANH MỤC BẢNG BIỂU

*Bảng 3.1.2 Cấu trúc và ý nghĩa bộ dữ liệu Weather*

# LỜI MỞ ĐẦU

1. **Mục tiêu của đề tài**

* Thời tiết đóng vai trò vô cùng quan trọng và không thể phủ nhận trong cuộc sống của con người. Nó ảnh hưởng đến mọi khía cạnh, từ sinh hoạt hàng ngày cho đến nền kinh tế và môi trường. Thời tiết có thể tạo nên những khoảnh khắc đáng nhớ và cũng có thể gây ra những tác động không mong muốn. Thời tiết quyết định cách chúng ta ăn mặc và tổ chức cuộc sống hàng ngày. Tùy thuộc vào nhiệt độ, mưa và ánh sáng mặt trời, chúng ta lựa chọn trang phục phù hợp và quyết định các hoạt động phù hợp trong ngày. Thời tiết đẹp và thoải mái giúp chúng ta cảm thấy hạnh phúc và năng động, trong khi thời tiết xấu có thể làm giảm tinh thần và đánh mất động lực.Trong lĩnh vực nông nghiệp, thời tiết là yếu tố quyết định thành công của mùa vụ. Lượng mưa, nhiệt độ và độ ẩm địa phương quyết định đến sự sinh trưởng của cây trồng và sự phát triển của vật nuôi.Ngoài ra, thời tiết cũng ảnh hưởng lớn đến hoạt động công nghiệp và sản xuất. Các ngành công nghiệp như năng lượng, xây dựng, vận chuyển cần phải điều chỉnh hoạt động của mình dựa trên điều kiện thời tiết. Thời tiết xấu có thể gây ra trong vận hành và ảnh hưởng đến hiệu suất và lợi nhuận của các doanh nghiệp.Thời tiết đóng vai trò quan trọng trong việc quản lý thiên tai và bảo vệ môi trường. Dự báo thời tiết giúp cảnh báo về các hiện tượng thiên tai như bão, lũ lụt, đất sạt, gió lớn... để người dân và cơ quan chức năng có thời gian chuẩn bị và ứng phó kịp thời, giảm thiểu thiệt hại và đảm bảo an toàn.Tóm lại, thời tiết đóng vai trò quan trọng và đa chiều trong cuộc sống của con người. Nó ảnh hưởng không chỉ đến sinh hoạt hàng ngày và nền kinh tế mà còn tới quản lý thiên tai và bảo vệ môi trường. Việc hiểu và đối mặt với thời tiết là cách để chúng ta tận dụng những cơ hội và đối mặt với thách thức một cách hiệu quả.
* Trong nền kinh tế thị trường, Phân tích dữ liệu và dự báo là vô cùng quan trọng trong việc hỗ trợ các nhà quản lý đưa ra phán đoán, quyết định đúng đắn dựa vào các dữ liệu đã thu thập được. Trên thế giới đã có nhiều tác giả đưa ra cách phân loại các phương pháp dự báo khác nhau. Các doanh nghiệp sử dụng những công cụ phần mềm và hệ thống để phân tích khối lượng dữ liệu lớn được thu thập trong một khoảng thời gian dài. Sau đó, phần mềm sẽ dự đoán nhu cầu cũng như xu hướng trong tương lai để giúp các công ty đưa ra những quyết định tài chính, tiếp thị và hoạt động chính xác hơn.
* Việc dự báo giúp các nhà quản lý doanh nghiệp tự tin hơn trong việc đối mặt với những thay đổi của thị trường, kiểm soát tốt hơn các hoạt động kinh doanh và đưa ra những quyết định chiến lược nhằm thúc đẩy sự tăng trưởng bền vững trong tương lai. Các doanh nghiệp có thể sử dụng dự báo vì nhiều mục đích khác nhau, ví dụ như tận dụng hiệu quả hơn các nguồn lực sẵn có, trực quan hóa các chỉ số hoạt động và hiệu suất kinh doanh, lựa chọn thời điểm phù hợp để ra mắt sản phẩm và dịch vụ mới, ước tính chi phí định kỳ, dự đoán các sự kiện quan trọng trong tương lai như doanh số bán hàng và doanh thu, cũng như đánh giá tính hiệu quả của các quyết định quản lý.

1. **Đối tượng và phạm vi nghiên cứu**

* Nghiên cứu tổng quan về phân tích dữ liệu và dự báo kinh doanh
* Xây dựng hệ thống phân tích dữ liệu và dự báo kinh doanh

1. **Kết cấu của đề tài**

Đề tài được tổ chức gồm phần mở đầu, 4 chương nội dung và phần kết luận...

* Mở đầu
* **Chương 1**: Tổng quan về đề tài áp dụng mô hình học máy trong phân tích dữ liệu dự thời tiết và giới thiệu Công Ty TNHH phần mềm Việt Đà
* **Chương 2**: Cơ sở lí thuyết phân tích dữ liệu và dự báo
* **Chương 3**: Các phương pháp phân tích dữ liệu và dự báo
* **Chương 4**: Hệ thống phân tích dữ liệu và dự báo thời tiết
* Kết luận và hướng phát triển

1. **TỔNG QUAN VỀ ĐỀ TÀI VÀ CÔNG TY TNHH PHẦN MỀM VIỆT ĐÀ**

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

### 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

Thứ hai, lịch sử hình thành và phát triển

Công ty TNHH Phần Mềm Việt Đà được thành lập ngày 30 tháng 4 năm 2008 với chức năng chính là sản xuất và cung ứng phần mềm ứng dụng như: phần mềm kế toán, nhà hàng, khách sạn, Quản lý bệnh viện, Quản lý bán hàng, nhân sự cho thị trường Đà nẵng và các tỉnh lân cận. Cùng với sự phát triển của nền kinh tế nhu cầu sử dụng phần mềm ngày một tăng lên tập thể Phần Mềm Việt Đà đã tìm hiểu và phát triển thị trường các tỉnh phía bắc.

Ngày 28 tháng 3 năm 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 cũng đã tiến hành thay đổi diện mạo, chuyển văn phòng về địa chỉ: 41 Trần Văn Ơn, P. Hòa An, Q.Cẩm Lệ, Tp.Đà Nẵng. Đăng ký thương hiệu sản phẩm VIỆT ĐÀ để tạo điều kiện thuận lợi cho quá trình phát triển kinh doanh về sau.

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.

Với tiêu chí không ngừng “Đổi Mới, Sáng Tạo” trong cách nghĩ, cách làm tập thể Phần Mềm Việt Đà sẽ luôn phấn đấu trở thành đơn vị dẫn đầu về sự hài lòng của khách hàng, cung cấp các “Giải pháp tối ưu” đến từng khách hàng. Chính sách chăm sóc khách hàng luôn được quan tâm, tận tình hỗ trợ từ quy trình sử dụng phần mềm đến nghiệp vụ công việc.

Phần Mềm Việt Đà là một tập thể vững mạnh, đoàn kết, chuyên nghiệp luôn cống hiến cho khách hàng những sản phẩm/dịch vụ tốt nhất.

**\*Sơ đồ cơ cấu tổ chức**

Ảnh có chứa văn bản, biểu đồ, ảnh chụp màn hình, Kế hoạch

Mô tả được tạo tự động

*Hình 1.1 Sơ đồ cơ cấu tổ chức của Công Ty*

(Nguồn: Website công ty TNHH phần mềm Việt Đà)

Chức năng, nhiệm của sơ đồ cơ cấu tổ chức công ty:

Giám đốc: Điều hành hoạt động chung của công ty, xây dựng chiến lược và kế hoạch kinh doanh.

Phó Giám đốc: Hỗ trợ Giám đốc trong công tác quản lý và điều hành công ty.

Phòng Kế toán: Quản lý tài chính, lập và kiểm soát ngân sách, thực hiện các báo cáo tài chính.

Phòng Kinh doanh: Xây dựng chiến lược và kế hoạch kinh doanh, tìm kiếm khách hàng tiềm năng, ký kết hợp đồng.

Phòng Kỹ thuật: Phát triển sản phẩm phần mềm, tư vấn và hỗ trợ kỹ thuật cho khách hàng.

Phòng Hành chính Nhân sự: Quản lý nhân sự, tuyển dụng, đào tạo, lương thưởng cho nhân viên.

Phòng Marketing: Xây dựng chiến lược marketing, quảng bá sản phẩm đến khách hàng tiềm năng.

Phòng Chăm sóc khách hàng: Hỗ trợ và giải quyết các vấn đề của khách hàng liên quan đến sản phẩm và dịch vụ.

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

**Thứ nhất, tầm nhìn**

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.

Với tầm nhìn đó, Công ty đặt ra các mục tiêu chiến lược sau:

Một là, phát triển các sản phẩm phần mềm đa dạng, đáp ứng tốt nhất nhu cầu quản lý doanh nghiệp của khách hàng. Công ty sẽ tập trung nghiên cứu và phát triển các phần mềm quản lý cho các lĩnh vực chủ chốt của nền kinh tế như:

+ Phần mềm kế toán, tài chính doanh nghiệp;

+ Phần mềm quản lý sản xuất, chuỗi cung ứng;

+ Phần mềm quản lý bán hàng, quản lý khách hàng;

+ Phần mềm nhân sự, lương, quản trị nguồn nhân lực;

+ Phần mềm quản lý chuỗi cửa hàng, siêu thị;

* Đồng thời phát triển thêm các ứng dụng phần mềm theo yêu cầu riêng biệt của từng đơn vị, ngành nghề kinh doanh.
* Hai là, không ngừng cải tiến để nâng cao chất lượng sản phẩm, dịch vụ. Công ty cam kết cung cấp ra thị trường các sản phẩm phần mềm với chất lượng hoàn hảo nhất, đáp ứng tốt nhất nhu cầu khách hàng. Các sản phẩm phần mềm của Việt Đà sẽ tuân thủ hoàn toàn các quy chuẩn, quy định về CNTT của Nhà nước, đảm bảo tính bảo mật và an toàn thông tin cho người dùng. Công ty cũng không ngừng cải tiến chất lượng dịch vụ hỗ trợ, đào tạo và tư vấn cho khách hàng, nhằm mang lại sự hài lòng tuyệt đối.
* Ba là, xây dựng đội ngũ nhân lực có trình độ chuyên môn cao, làm chủ công nghệ. Công ty xác định nguồn nhân lực là yếu tố quyết định thành công lâu dài. Vì vậy, Việt Đà sẽ không ngừng đầu tư nâng cao trình độ đội ngũ, đặc biệt là đội ngũ kỹ sư phần mềm, cả về chuyên môn, nghiệp vụ và khả năng làm chủ công nghệ. Công ty cũng xây dựng môi trường làm việc chuyên nghiệp, thu hút nhân tài và thúc đẩy sự sáng tạo, dám nghĩ dám làm trong đội ngũ.
* Bốn là, không ngừng mở rộng thị trường trong cả nước. Hiện Công ty đã có mặt tại hầu hết các tỉnh thành phía Nam, và đang đẩy mạnh mở rộng ra các thị trường miền Trung, miền Bắc. Việt Đà đặt mục tiêu sẽ có mặt tại 63 tỉnh thành trên cả nước trong vòng 5 năm tới, cung cấp sản phẩm phần mềm cho hàng trăm ngàn doanh nghiệp trên khắp Việt Nam. Công ty cũng sẽ mở rộng thị trường ra các nước ASEAN trong tương lai.
* Năm là, liên tục đổi mới sáng tạo, ứng dụng công nghệ 4.0. Công ty nhận thức sự phát triển không ngừng của cuộc Cách mạng Công nghiệp 4.0 sẽ tạo ra những thay đổi lớn về nhu cầu và cách thức quản lý kinh doanh của doanh nghiệp. Vì vậy, Việt Đà cam kết sẽ liên tục đổi mới sáng tạo, tích hợp những công nghệ mới nhất như AI, Big Data, IoT, Blockchain vào các sản phẩm phần mềm nhằm tối ưu hóa giải pháp cho khách hàng.
* Để hiện thực hóa tầm nhìn và các mục tiêu chiến lược trên đây, Công ty TNHH Phần mềm Việt Đà sẽ tập trung nguồn lực vào các nhiệm vụ then chốt sau: Đầu tư nghiên cứu phát triển sản phẩm, tuyển dụng nhân sự, xây dựng quy trình phát triển phần mềm chuẩn quốc tế.Xây dựng văn hóa chất lượng, coi chất lượng sản phẩm và dịch vụ là yếu tố sống còn của Công ty. Xây dựng chiến lược nhân sự dài hạn, thu hút nhân tài, đào tạo và phát triển nguồn nhân lực chất lượng cao. Mở rộng thị trường một cách bài bản đến từng vùng miền, ngành nghề trên toàn quốc. Tiên phong ứng dụng công nghệ mới, không ngừng cải tiến để nâng cao giá trị của sản phẩm và dịch vụ.Xây dựng hệ thống quản trị Công ty theo chuẩn mực quốc tế.
* Với tầm nhìn và chiến lược phát triển như trên, Công ty TNHH Phần mềm Việt Đà tin tưởng sẽ đạt được mục tiêu trở thành công ty phần mềm hàng đầu tại Việt Nam, vươn tầm ra khu vực, góp phần nâng cao năng lực ứng dụng CNTT cho các doanh nghiệp, từ đó đóng góp vào sự phát triển chung của nền kinh tế.

**Thứ hai, sứ mệnh**

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.

Với sứ mệnh đó, Việt Đà cam kết thực hiện các giá trị cốt lõi sau:

Một là, luôn đặt lợi ích của khách hàng lên trên hết. Công ty luôn xem khách hàng là trung tâm, lấy sự hài lòng của khách hàng làm thước đo cho mọi hoạt động của mình. Việt Đà cam kết mang đến cho khách hàng những giải pháp tối ưu nhất, vượt trội nhất, đem lại giá trị và lợi ích thiết thực nhất cho hoạt động kinh doanh của khách hàng.

Hai là, luôn đổi mới sáng tạo để phát triển các sản phẩm, dịch vụ CNTT chất lượng cao. Với tinh thần dám nghĩ, dám làm, Việt Đà không ngừng nghiên cứu, ứng dụng khoa học công nghệ tiên tiến nhất để phát triển các giải pháp CNTT vượt trội. Công ty cũng liên tục cải tiến và nâng cấp các sản phẩm, dịch vụ hiện có để đáp ứng tốt nhất nhu cầu đa dạng và luôn thay đổi của thị trường.

Ba là, xây dựng nguồn nhân lực chất lượng cao. Việt Đà xác định con người chính là nhân tố quyết định sự phát triển bền vững của Công ty. Vì vậy, Việt Đà không ngừng đào tạo, nâng cao tay nghề cho đội ngũ nhân viên, xây dựng đội ngũ chuyên gia CNTT giỏi chuyên môn, giàu kinh nghiệm, đáp ứng yêu cầu ngày càng cao của thị trường.

Bốn là, phát triển bền vững cùng sự thịnh vượng chung. Việt Đà luôn đảm bảo sự cân bằng giữa lợi ích của doanh nghiệp, của khách hàng, của nhân viên và của cộng đồng. Công ty kinh doanh dựa trên các giá trị nhân văn, đạo đức, trách nhiệm xã hội cao. Đây chính là nền tảng cho sự phát triển bền vững, mang lại lợi ích lâu dài cho tất cả các bên.

Năm là, phục vụ sự nghiệp công nghiệp hóa, hiện đại hóa đất nước. Việt Đà nhận thức CNTT chính là động lực then chốt thúc đẩy quá trình công nghiệp hóa, hiện đại hóa. Vì vậy, Việt Đà luôn phấn đấu để cung cấp ra thị trường những sản phẩm CNTT chất lượng cao, góp phần nâng cao năng lực quản lý và năng suất lao động cho các doanh nghiệp, từ đó góp phần vào sự nghiệp công nghiệp hóa – hiện đại hóa đất nước.

Để thực hiện sứ mệnh và các giá trị cốt lõi nói trên, Việt Đà sẽ tập trung vào các chiến lược hành động sau:

* Liên tục nghiên cứu, ứng dụng công nghệ mới để phát triển các sản phẩm, giải pháp CNTT chất lượng cao, đáp ứng linh hoạt nhu cầu khách hàng.
* Đầu tư mạnh mẽ để nâng cao chất lượng nguồn nhân lực, xây dựng đội ngũ kỹ sư phần mềm chuyên nghiệp hàng đầu.
* Xây dựng chiến lược và mở rộng thị trường một cách hệ thống, bài bản với tầm nhìn dài hạn.
* Triển khai các hoạt động hướng về cộng đồng, thể hiện trách nhiệm doanh nghiệp với xã hội.
* Xây dựng văn hóa doanh nghiệp chuyên nghiệp, coi trọng đạo đức kinh doanh và sự thịnh vượng chung.
* Với đội ngũ nhân viên tâm huyết, nhiệt huyết, giàu chuyên môn và kinh nghiệm, cùng hệ thống quản trị chặt chẽ theo chuẩn mực quốc tế, Công ty TNHH Phần mềm Việt Đà tin tưởng sẽ hoàn thành sứ mệnh cao đẹp của mình, góp phần đưa CNTT Việt Nam hội nhập sâu rộng với thế giới, vươn tầm đẳng cấp quốc tế.

**Thứ ba, ngành nghề kinh doanh của công ty**

* 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 kế toán được thiết kế phù hợp với các doanh nghiệp có quy mô vừa và nhỏ trong nhiều lĩnh vực khác nhau như thương mại, dịch vụ, sản xuất, giúp quản lý có được bức tranh tổng thể về tình hình tài chính của doanh nghiệp. Hỗ trợ kế toán tổng hợp, kế toán chi tiết các nghiệp vụ phát sinh hàng ngày như: mua bán hàng hóa, công nợ phải thu phải trả, tính giá thành, lương nhân viên. Cung cấp các báo cáo tổng hợp và chi tiết về tình hình tài chính doanh nghiệp như bảng cân đối tài khoản, báo cáo kết quả hoạt động kinh doanh, báo cáo lưu chuyển tiền tệ,... để phục vụ công tác quản trị và ra quyết định. Tuân thủ chuẩn mực kế toán Việt Nam, hỗ trợ kê khai thuế điện tử và các quy định pháp lý hiện hành.

+ Phần mềm nhà hàng, khách sạn: Quản lý các hoạt động bán hàng như đặt bàn, order món ăn, thanh toán tại quầy bar và bàn ăn. Quản lý các dịch vụ phòng và các dịch vụ khác của khách sạn như giặt ủi, spa, vận chuyển. Quản lý kho, hàng hóa, thực đơn, giá cả và chi phí. Định mức và tính giá thành món ăn, giá phòng. Thống kê và báo cáo hoạt động kinh doanh nhà hàng/khách sạn, doanh thu và chi phí một cách chi tiết. Hỗ trợ việc điều hành và ra quyết định kinh doanh hiệu quả.

+ Phần mềm quản lý bán hàng: Quản lý danh mục khách hàng, sản phẩm, kho hàng. Theo dõi và quản lý các đơn hàng bán ra, đơn đặt hàng của khách, lịch sử mua hàng. Quản lý chiến dịch tiếp thị, chương trình khuyến mãi. Thống kê báo cáo tổng hợp về hoạt động kinh doanh, bán hàng để đánh giá hiệu quả. Hỗ trợ công tác quản trị bán hàng và marketing của doanh nghiệp.

+ Phần mềm quản lý kho, hàng hóa: Quản lý danh mục hàng hóa, nhà cung cấp, khách hàng. Theo dõi tồn kho, xuất nhập hàng hóa, phiếu xuất kho, phiếu nhập kho. Quản lý đơn đặt hàng, đơn hàng nhập/xuất, theo dõi đơn hàng đã giao/ nhận. Báo cáo tồn kho, tình hình nhập xuất hàng, công nợ phải thu phải trả. Hỗ trợ công tác quản lý kho, vật tư, hàng hóa và điều phối chuỗi cung ứng.

+ Phần mềm quản trị nhân lực: Quản lý thông tin nhân sự của công ty như hồ sơ cá nhân, quá trình công tác. Lập kế hoạch nhân lực, tuyển dụng, đào tạo và phát triển nhân viên. Tính toán, quản lý lương, thưởng, các khoản phụ cấp và trích nộp BHXH. Thống kê, phân tích và báo cáo về tình hình sử dụng lao động của công ty. Hỗ trợ công tác nhân sự và quản trị nguồn nhân lực của doanh nghiệp.

Ngoài cung cấp phần mềm, Công ty TNHH Phần mềm Việt Đà còn cung cấp các dịch vụ tư vấn và hỗ trợ kèm theo như: Tư vấn lựa chọn và cài đặt phần mềm phù hợp với nhu cầu sử dụng của khách hàng; Đào tạo và hướng dẫn chi tiết cách sử dụng phần mềm sau khi cài đặt; Hỗ trợ kỹ thuật về các lỗi phần mềm có thể gặp phải trong quá trình sử dụng; Cập nhật phiên bản mới khi phần mềm có bản nâng cấp; Tư vấn về nghiệp vụ, quy trình liên quan đến lĩnh vực mà phần mềm hỗ trợ; Hỗ trợ di chuyển dữ liệu khi khách hàng chuyển đổi phần mềm mới; Dịch vụ sao lưu dữ liệu định kỳ cho khách hàng.

Với đội ngũ kỹ sư phần mềm giàu kinh nghiệm và am hiểu nghiệp vụ, Công ty TNHH Phần mềm Việt Đà có thể phát triển và cung cấp các ứng dụng phần mềm theo yêu cầu riêng của từng khách hàng. Các phần mềm được xây dựng dựa trên nền tảng công nghệ hiện đại, đáp ứng các tiêu chí về tính năng, hiệu năng và bảo mật.

Với hơn 10 năm kinh nghiệm trong lĩnh vực phần mềm doanh nghiệp, Công ty TNHH Phần mềm Việt Đà đã xây dựng được uy tín với khách hàng thông qua các sản phẩm và dịch vụ có chất lượng, đem lại hiệu quả thiết thực cho doanh nghiệp. Công ty hiện đã cung cấp phần mềm cho hơn 5.000 doanh nghiệp trên khắp cả nước thuộc nhiều lĩnh vực và ngành nghề khác nhau. Một số khách hàng tiêu biểu có thể kể đến:

Khách sạn Sheraton Đà Nẵng;

Khách sạn Vinpearl Nam Hội An;

Tập đoàn Sovico;

Công ty Cổ phần Bibica;

Trung tâm thương mại Go Mall Đà Nẵng;

FPT Software Miền Trung;

Công ty TNHH Giày Kim Việt;

Công ty CP Dược phẩm TW25;

Ngân hàng NN&PTNT chi nhánh Quảng Nam;

Sở Tài chính tỉnh Quảng Ngãi;

Bệnh viện Đa khoa Đà Nẵng.

Với hệ thống phân phối rộng khắp trên toàn quốc cùng đội ngũ nhân viên tư vấn giỏi, Công ty TNHH Phần mềm Việt Đà có thể đáp ứng nhu cầu đa dạng của khách hàng về phần mềm quản lý doanh nghiệp. Khách hàng của Công ty không chỉ đến từ các tỉnh thành phía Nam mà còn ở các khu vực miền Trung, miền Bắc. Điều này cho thấy uy tín và năng lực cung cấp phần mềm, dịch vụ CNTT của Công ty đã được khẳng định trên phạm vi toàn quốc.

## Tổng quan về đề tài

* + 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. Trong lĩnh vực dự báo thời tiết, mô hình học máy có thể phân tích hàng loạt dữ liệu phức tạp như nhiệt độ, áp suất không khí, độ ẩm, và gió để tạo ra các dự đoán thời tiết chính xác hơn. 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: Dữ liệu thời tiết thường có độ biến đổi lớn và không đồng đều. Các giá trị dữ liệu có thể bị thiếu hoặc bị nhiễu, ảnh hưởng bởi nhiều yếu tố khác nhau. Điều này đặt ra thách thức trong việc xây dựng mô hình học máy đáng tin cậy từ dữ liệu không chắc chắn.
* Độ phức tạp của mô hình: Mô hình học máy có thể có độ phức tạp khác nhau, từ các mô hình đơn giản như hồi quy tới các mạng nơ-ron sâu. Việc xác định mô hình phù hợp và cần thiết để đạt được độ chính xác cao trong dự đoán thời tiết là một vấn đề quan trọng.
* Cơ sở dữ liệu lớn và xử lý dữ liệu: Dữ liệu thời tiết có thể rất lớn và phức tạp. Việc thu thập, lưu trữ và xử lý dữ liệu đòi hỏi khả năng tính toán mạnh mẽ và quản lý hiệu quả của hệ thống.
* Điều kiện biên và biến đổi thời tiết cực đoan: Mô hình học máy có thể gặp khó khăn khi phải dự đoán các tình huống thời tiết cực đoan hoặc biến đổi đột ngột. Các tình huống như cơn bão, sự biến đổi khí hậu, hay thay đổi thời tiết đột ngột đều đòi hỏi mô hình có khả năng ứng phó tốt.
* Mặc dù có những thách thức này, việc áp dụng mô hình học máy trong phân tích và dự đoán dữ liệu thời tiết hứa hẹn mang lại những tiềm năng lớn để cải thiện khả năng dự báo, nâng cao an toàn và tối ưu hóa các hoạt động trong nhiều lĩnh vực.
  + 1. 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: Tiến xa hơn trong nghiên cứu và phát triển các mô hình học máy để dự đoán thời tiết. Sử dụng các kiến thức từ lĩnh vực trí tuệ nhân tạo và học sâu để xây dựng các mạng nơ-ron phức tạp hơn, có khả năng học cấu trúc và mối quan hệ phức tạp trong dữ liệu thời tiết.
  + Kết hợp dữ liệu đa nguồn: Kết hợp dữ liệu thời tiết từ nhiều nguồn khác nhau như cảm biến đất, không gian và biển để tạo ra dự đoán thời tiết chính xác hơn và bao quát hơn. Điều này có thể cải thiện khả năng dự đoán trong các điều kiện đặc biệt và biến đổi khí hậu.
  + Xử lý dữ liệu thời tiết không chắc chắn: Phát triển phương pháp xử lý dữ liệu thời tiết không chắc chắn và nhiễu. Điều này có thể bao gồm việc sử dụng các mô hình xác suất để ước tính mức độ không chắc chắn của dự đoán thời tiết và cung cấp thông tin về độ tin cậy của dự đoán.
  + Dự báo thời tiết cực đoan: Tập trung vào khả năng dự đoán các tình huống thời tiết cực đoan như cơn bão, lũ lụt, hay sự biến đổi khí hậu. Xây dựng mô hình có khả năng ứng phó với các biến đổi đột ngột và tình huống khẩn cấp.
  + Phân tích ứng dụng thực tế: Áp dụng mô hình học máy trong các lĩnh vực thực tế như quản lý giao thông, nông nghiệp, năng lượng, và du lịch. Đánh giá hiệu suất của mô hình trong các tình huống thực tế và đo lường tác động thực sự của việc cải thiện dự đoán thời tiết.
  + Khám phá phân tích dữ liệu sâu hơn: Sử dụng các phương pháp phân tích dữ liệu sâu hơn để hiểu rõ hơn về mối quan hệ giữa các yếu tố thời tiết, nhưng cũng các yếu tố nhân tạo và môi trường tác động lên thời tiết.

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.

1. **CƠ SỞ LÝ THUYẾT PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO THỜI TIẾT**

## Giới thiệu về Data Analyst

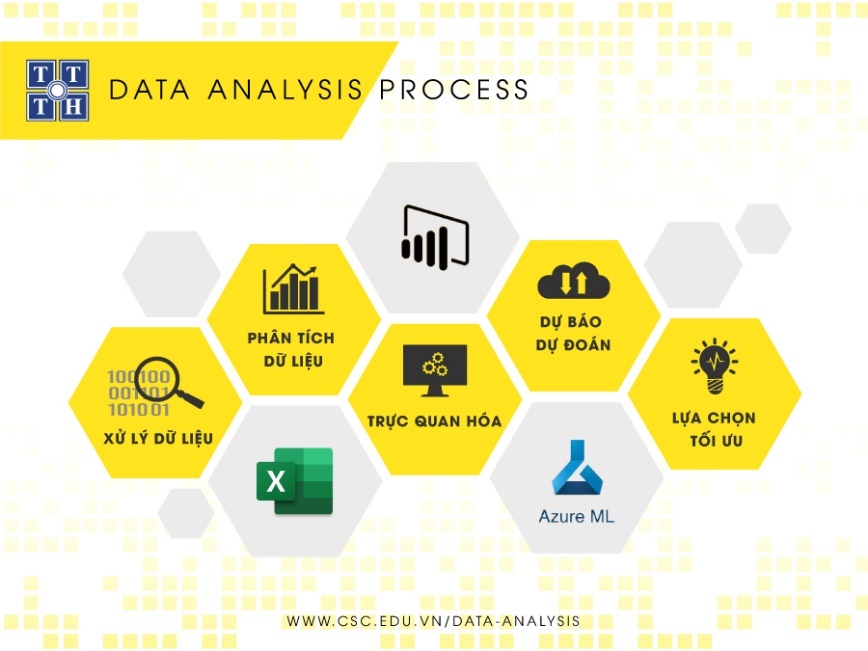
* Vị trí công việc Data Analyst đang trở nên ngày càng quan trọng và phổ biến trong xã hội hiện nay. Sự phát triển nhanh chóng của công nghệ thông tin và khai phá dữ liệu đã tạo ra một lượng lớn thông tin và dữ liệu có giá trị. Data Analyst đóng vai trò quan trọng trong việc biến các dữ liệu này thành thông tin hữu ích và ý nghĩa cho các doanh nghiệp, tổ chức, và xã hội. Trong thời đại số hóa và kỷ nguyên công nghệ 4.0, việc thu thập và phân tích dữ liệu đã trở thành một phần không thể thiếu trong quyết định kinh doanh và quản lý. 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. Họ xây dựng các mô hình dự đoán, phân tích thống kê và đánh giá tác động của các quyết định kinh doanh. Công việc của Data Analyst không chỉ giới hạn trong doanh nghiệp, mà còn ở nhiều lĩnh vực khác nhau của xã hội. 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. Các chuyên gia phân tích dữ liệu không chỉ cần có kiến thức vững về khoa học dữ liệu và phân tích, mà còn cần có khả năng diễn giải và giao tiếp kết quả phân tích một cách rõ ràng và dễ hiểu cho các đối tượng khác nhau. Tóm lại, Data Analyst đóng vai trò quan trọng và cần thiết trong xã hội hiện nay. Việc khai thác và tận dụng dữ liệu để hiểu rõ hơn về thế giới xung quanh và tạo ra giá trị là một trong những xu hướng chính đang thúc đẩy sự phát triển và tiến bộ của xã hội.

### Khái niệm:

* 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. Vai trò của Data Analyst là giúp các tổ chức và doanh nghiệp hiểu rõ hơn về dữ liệu mà họ sở hữu và từ đó đưa ra quyết định thông minh và hiệu quả.
* 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. Họ sử dụng các công cụ và phần mềm phân tích dữ liệu như Microsoft Excel, Python, , SQL, và các công cụ khai phá dữ liệu và học máy để thực hiện công việc phân tích.
* 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. Họ xây dựng các biểu đồ, báo cáo, và các mô hình dự đoán để hỗ trợ việc ra quyết định kinh doanh, dự báo xu hướng và định hướng phát triển tương lai.
* 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. Các Data Analyst có thể làm việc trong nhiều lĩnh vực khác nhau như tài chính, marketing, y tế, giáo dục và các ngành công nghiệp khác, hỗ trợ quyết định và tối ưu hóa hoạt động của tổ chức và doanh nghiệp.

### 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: Data Analyst tập trung vào việc thu thập dữ liệu từ nhiều nguồn khác nhau trong doanh nghiệp, bao gồm dữ liệu từ hệ thống thông tin, cơ sở dữ liệu, tệp Excel, trang web, và các nguồn bên ngoài khác.
* Xử lý và làm sạch dữ liệu: Trước khi tiến hành phân tích, Data Analyst cần xử lý và làm sạch dữ liệu để loại bỏ các sai sót, dữ liệu trùng lặp, hoặc dữ liệu thiếu sót. Điều này đảm bảo rằng dữ liệu đang được sử dụng là chính xác và đáng tin cậy.
* Phân tích dữ liệu: Sau khi làm sạch dữ liệu, Data Analyst sử dụng các kỹ thuật phân tích thống kê và học máy để khám phá thông tin tiềm ẩn trong dữ liệu. Họ tạo ra các biểu đồ, báo cáo, và các mô hình dự đoán để trình bày kết quả phân tích.
* Đưa ra insights và giải pháp: Data Analyst đưa ra insights và giải pháp dựa trên kết quả phân tích để hỗ trợ quyết định kinh doanh. Những thông tin này giúp doanh nghiệp hiểu rõ hơn về thị trường, khách hàng, sản phẩm và hoạt động kinh doanh
* Theo dõi và đánh giá hiệu suất: Data Analyst cũng đóng vai trò trong việc theo dõi hiệu suất kinh doanh và đánh giá hiệu quả của các chiến lược và chương trình trong doanh nghiệp. Họ giúp xác định các chỉ số hiệu suất và theo dõi sự thay đổi của chúng theo thời gian.
* Dự báo và kế hoạch tương lai: Data Analyst sử dụng dữ liệu hiện có để dự báo xu hướng và phân tích các kịch bản tương lai. Những dự báo này hỗ trợ doanh nghiệp trong việc lập kế hoạch và đưa ra quyết định dựa trên tương lai dự kiến.
* 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.



*Hình 2.1.2 Các công việc của data analyst*

## 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. Trong hoạt động kinh doanh, dữ liệu thời tiết đóng góp một số lợi ích quan trọng như sau: Chiến lược tiếp thị: Các doanh nghiệp có thể sử dụng dữ liệu thời tiết để xác định xu hướng tiêu thụ và thích ứng chiến lược tiếp thị của họ. Ví dụ, các cửa hàng bán lẻ có thể điều chỉnh chiến lược quảng cáo và giảm giá hàng hóa dựa trên dự báo thời tiết, như quảng cáo đồ đạc ngoài trời trong những ngày nắng đẹp hoặc giảm giá áo khoác trong những ngày lạnh. Quản lý chuỗi cung ứng: Dữ liệu thời tiết có thể giúp các doanh nghiệp dự đoán và ứng phó với các biến đổi trong nhu cầu sản phẩm. Ví dụ, các nhà sản xuất thực phẩm có thể dự đoán nhu cầu của các sản phẩm mùa trong các thời điểm cụ thể và điều chỉnh quy trình sản xuất và chuỗi cung ứng của họ để đáp ứng nhu cầu này. Quản lý rủi ro: Dữ liệu thời tiết cung cấp thông tin quan trọng về các điều kiện thời tiết bất thường như bão, lũ lụt, tuyết lở, và hạn hán. Các doanh nghiệp có thể sử dụng thông tin này để lập kế hoạch và đối phó với những tác động tiêu cực của thời tiết, giảm thiểu rủi ro và thiệt hại. Dự báo năng suất nông nghiệp: Dữ liệu thời tiết có thể giúp các doanh nghiệp nông nghiệp dự báo năng suất và chất lượng mùa vụ. Việc hiểu rõ về điều kiện thời tiết có thể giúp cho việc lựa chọn loại cây trồng, thời điểm trồng, và quy trình chăm sóc phù hợp, từ đó tối ưu hóa năng suất và lợi nhuận. 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. Việc sử dụng thông tin từ dữ liệu thời tiết giúp cải thiện chiến lược tiếp thị, quản lý chuỗi cung ứng, đối phó với rủi ro và tối ưu hóa hoạt động sản xuất và kinh doanh.

### 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 được thu thập thông qua các trạm quan trắc thời tiết, các cảm biến trên vệ tinh, các máy bay và các thiết bị phát sóng đặc biệt trên mặt đất. Các thông tin này sau đó được ghi lại và cập nhật thường xuyên để cung cấp các dữ liệu thời tiết chính xác và đáng tin cậy. 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. Trong khi đó, 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 đời :

* Dữ liệu thời tiết đóng vai trò vô cùng quan trọng trong đời sống của chúng ta, ảnh hưởng đến nhiều khía cạnh khác nhau từ việc lựa chọn trang phục hàng ngày cho đến quyết định kinh doanh lớn. Tầm quan trọng của dữ liệu thời tiết không chỉ dừng lại ở việc cung cấp thông tin về thời tiết hiện tại, mà còn mở ra những cơ hội đáng kể trong việc hiểu rõ và ứng phó với biến đổi khí hậu.
* Dữ liệu thời tiết giúp chúng ta chuẩn bị tốt hơn cho các hoạt động hàng ngày. Dự báo thời tiết giúp bạn biết được mọi thay đổi về nhiệt độ, mưa và gió, từ đó lựa chọn trang phục phù hợp và chuẩn bị kế hoạch cho các hoạt động ngoài trời.
* Trong lĩnh vực nông nghiệp, dữ liệu thời tiết là một yếu tố quyết định trong việc quản lý vụ mùa và sản xuất cây trồng. Việc theo dõi thông tin về nhiệt độ, mưa và độ ẩm giúp nông dân đưa ra quyết định đúng đắn về thời điểm gieo trồng, tưới tiêu và thu hoạch, từ đó tối ưu hóa sản lượng và chất lượng.
* Trong lĩnh vực kinh doanh, dữ liệu thời tiết có thể giúp doanh nghiệp tối ưu hóa hoạt động và định hướng chiến lược. Tích hợp dữ liệu thời tiết vào quyết định kinh doanh giúp doanh nghiệp quản lý nguồn lực, điều chỉnh chiến lược tiếp thị và quản lý chuỗi cung ứng một cách hiệu quả hơn.
* Ngoài ra, dữ liệu thời tiết còn có vai trò quan trọng trong việc nghiên cứu về biến đổi khí hậu và ảnh hưởng của nó đến môi trường và cuộc sống con người. Dữ liệu thời tiết cung cấp dữ liệu cho các nhà khoa học để phân tích và đưa ra dự đoán về tương lai, từ đó hỗ trợ trong việc xây dựng các chính sách bảo vệ môi trường và quản lý tài nguyên tự nhiên.
* Tóm lại, dữ liệu thời tiết có tầm quan trọng không thể bàn cãi trong cuộc sống hàng ngày của chúng ta. Nó ảnh hưởng đến cách chúng ta sống, làm việc và quản lý tài nguyên, đồng thời giúp chúng ta hiểu rõ hơn về thế giới xung quanh và tạo điều kiện cho sự phát triển bền vững.

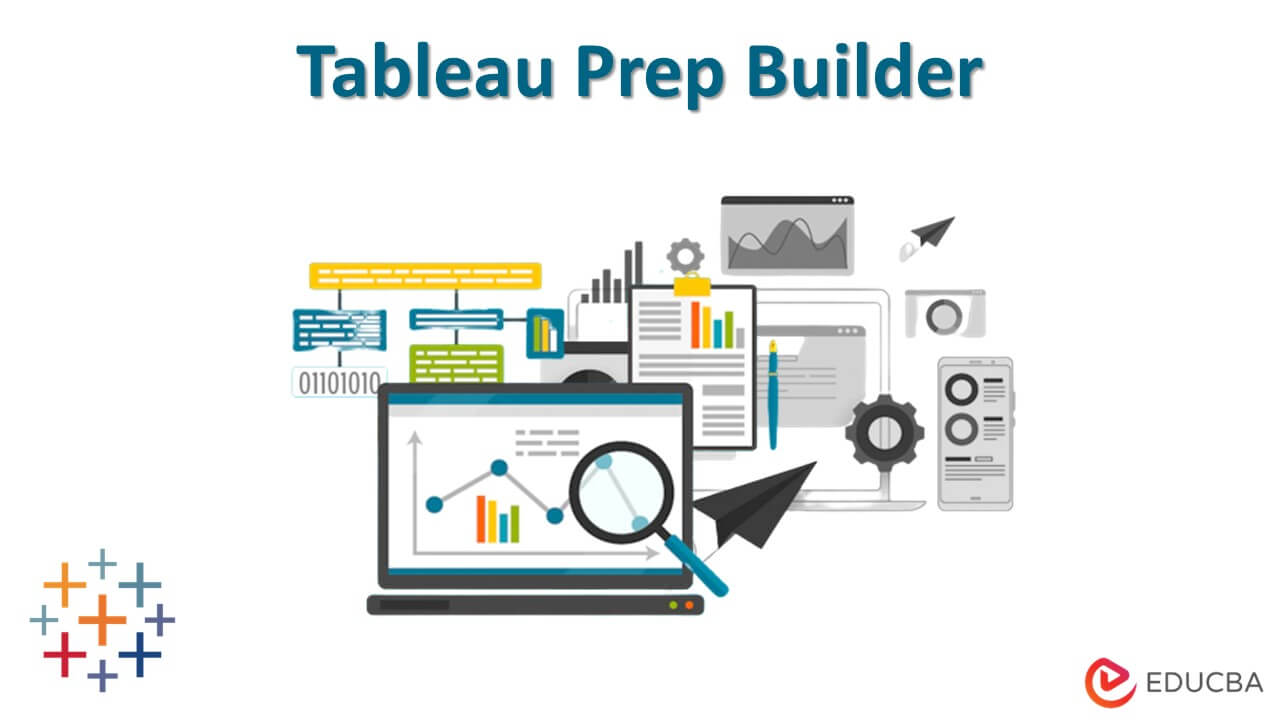
### 2.2.3 Lợi ích và ứng dụng của phân tích dữ liệu thời tiết trong đời sống:

Phân tích dữ liệu thời tiết đem lại lợi ích và ứng dụng đa dạng trong đời sống, từ việc chuẩn bị cho ngày mới đến quyết định kinh doanh và bảo vệ môi trường. Dưới đây là một số lợi ích và ứng dụng quan trọng của việc phân tích dữ liệu thời tiết:

* + - * Dự báo thời tiết chính xác: Phân tích dữ liệu thời tiết giúp tạo ra các dự báo thời tiết chính xác hơn. Điều này giúp mọi người chuẩn bị cho các biến đổi thời tiết, từ việc lựa chọn trang phục hàng ngày cho đến kế hoạch đi lại và hoạt động ngoại trời.
      * Quản lý nguồn lực: Dữ liệu thời tiết giúp quản lý nguồn lực hiệu quả. Nông nghiệp có thể dự đoán thời tiết để quản lý sản xuất cây trồng, và doanh nghiệp cũng có thể tối ưu hóa hoạt động dựa trên thông tin thời tiết.
* Chiến lược tiếp thị: Dữ liệu thời tiết ảnh hưởng đến hành vi mua sắm của khách hàng. Các doanh nghiệp có thể tận dụng thông tin này để điều chỉnh chiến lược tiếp thị và quảng cáo phù hợp với điều kiện thời tiết.
* Quản lý chuỗi cung ứng: Dữ liệu thời tiết giúp dự đoán tình hình chuỗi cung ứng. Các doanh nghiệp có thể phản ứng nhanh với thay đổi thời tiết, đảm bảo nguồn cung cấp và giảm thiểu rủi ro.
* Đối phó với thay đổi khí hậu: Dữ liệu thời tiết cung cấp thông tin quan trọng cho việc nghiên cứu và đối phó với thay đổi khí hậu. Điều này giúp xây dựng các chính sách bảo vệ môi trường và phát triển bền vững.
* An toàn và quản lý rủi ro: Dữ liệu thời tiết giúp đảm bảo an toàn cho con người và tài sản. Điều này đặc biệt quan trọng trong việc đối phó với thời tiết cực đoan như cơn bão, lũ lụt hay nhiệt độ cao.
* Phân tích dữ liệu thời tiết mang đến nhiều lợi ích và ứng dụng quan trọng trong đời sống. Từ việc cải thiện dự báo thời tiết cho đến tối ưu hóa hoạt động kinh doanh và bảo vệ môi trường, dữ liệu thời tiết đóng vai trò quan trọng trong việc tối ưu hóa cuộc sống hàng ngày và xây dựng tương lai bền vững.

2.3 Công cụ xử lý dữ liệu Tableau Prep Builder

2.3.1. Khái niệm Tableau Prep Builder:



Hình 2.3.1 Tableau Prep Builder

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: Kết nối và nhập dữ liệu từ nhiều nguồn khác nhau như cơ sở dữ liệu, tệp Excel, tệp văn bản, và nhiều nguồn dữ liệu trực tuyến khác.

Biến đổi dữ liệu: Thực hiện các biến đổi trên dữ liệu như ghép bảng, loại bỏ cột trống, chuyển đổi kiểu dữ liệu, tách cột thành nhiều cột, và nhiều hoạt động biến đổi dữ liệu khác.

Làm sạch dữ liệu: Lọc và loại bỏ dữ liệu không hợp lệ, giá trị trùng lặp, và các giá trị không chính xác hoặc thiếu sót.

Tích hợp dữ liệu: Kết hợp dữ liệu từ các nguồn khác nhau thành một bộ dữ liệu chung để tiến hành phân tích và trực quan hóa sau đó.

Tableau Prep Builder có giao diện dễ sử dụng, giúp người dùng thao tác một cách trực quan và tương tác với dữ liệu một cách linh hoạt. Sau khi dữ liệu đã được chuẩn bị sẵn sàng, người dùng có thể xuất dữ liệu vào Tableau Desktop để tiến hành phân tích và trực quan hóa dữ liệu một cách chuyên nghiệp và hiệu quả.

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: Tableau Prep Builder được thiết kế với giao diện trực quan, dễ sử dụng và thân thiện với người dùng. Người dùng có thể thực hiện các biến đổi dữ liệu và làm sạch dữ liệu thông qua các thao tác kéo và thả, giúp giảm thời gian và công sức trong quá trình chuẩn bị dữ liệu.
* Chuẩn bị dữ liệu đa nguồn: Tableau Prep Builder hỗ trợ kết nối và chuẩn bị dữ liệu từ nhiều nguồn khác nhau như cơ sở dữ liệu, tệp Excel, tệp văn bản, và nhiều nguồn dữ liệu trực tuyến khác. Điều này giúp người dùng dễ dàng thao tác với các nguồn dữ liệu phức tạp và đa dạng.
* Xem trước dữ liệu: Trước khi áp dụng các biến đổi và làm sạch dữ liệu, người dùng có thể xem trước dữ liệu để hiểu rõ cấu trúc và nội dung của dữ liệu. Điều này giúp người dùng đưa ra quyết định chính xác về các biến đổi và làm sạch dữ liệu cần thực hiện.
* Tự động phát hiện sự không phù hợp: Tableau Prep Builder có khả năng tự động phát hiện và gợi ý các vấn đề trong dữ liệu như dữ liệu trùng lặp, giá trị không hợp lệ, và các giá trị thiếu sót. Điều này giúp người dùng dễ dàng và nhanh chóng sửa chữa các vấn đề này trong quá trình chuẩn bị dữ liệu.
* Lịch sử thay đổi: Tableau Prep Builder ghi lại lịch sử thay đổi của dữ liệu, cho phép người dùng theo dõi quá trình chuẩn bị dữ liệu và quay lại các bước trước đó nếu cần thiết. Điều này giúp đảm bảo tính chính xác và đáng tin cậy của dữ liệu sau khi đã được chuẩn bị.
* Tích hợp với Tableau Desktop: Tableau Prep Builder tích hợp hoàn hảo với Tableau Desktop, cho phép người dùng chuyển dữ liệu đã được chuẩn bị từ Prep Builder sang Tableau Desktop để tiến hành phân tích và trực quan hóa dữ liệu.
* 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:



Hình 2.4.1 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: Tableau cho phép người dùng trực quan hóa dữ liệu thông qua các biểu đồ, biểu đồ hình ảnh, bản đồ, bảng, biểu đồ cây, v.v. Điều này giúp người dùng dễ dàng nhìn thấy và hiểu rõ hơn về mô hình, mẫu và xu hướng trong dữ liệu.
* Kết nối đa nguồn: Tableau hỗ trợ kết nối và thao tác dữ liệu từ nhiều nguồn khác nhau, bao gồm cơ sở dữ liệu, tệp Excel, tệp văn bản, dữ liệu trực tuyến và nhiều nguồn dữ liệu khác.
* Tích hợp dữ liệu: Tableau cho phép người dùng kết hợp dữ liệu từ các nguồn khác nhau để tạo thành một bộ dữ liệu chung để tiến hành phân tích và trực quan hóa dữ liệu.
* Tự động cập nhật: Tableau tự động cập nhật dữ liệu khi có sự thay đổi, giúp người dùng luôn có dữ liệu mới nhất để phân tích và trực quan hóa.Tương tác và trình diễn dữ liệu: Tableau cho phép người dùng tương tác với dữ liệu và trình diễn dữ liệu một cách chuyên nghiệp thông qua các báo cáo tương tác và bảng điều khiển.
* Khả năng chia sẻ và xuất dữ liệu: Tableau cho phép người dùng chia sẻ báo cáo và biểu đồ với người khác bằng cách xuất dữ liệu dưới dạng hình ảnh, PDF, hoặc các định dạng tập tin khác.
* 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:

2.4.3.1.Q&A Dashboard

Q&A Dashboard là một loại bảng điều khiển được thiết kế để hiển thị và quản lý thông tin liên quan đến việc trả lời câu hỏi (Q&A - Questions and Answers). Mục tiêu chính của Q&A Dashboard là cung cấp một cách tiếp cận tổ chức và dễ quản lý cho việc xử lý các câu hỏi từ người dùng, khách hàng hoặc nhân viên.**.** Nó có thể tích hợp các tính năng như: Tổ chức thông tin, Tìm kiếm và truy xuất, Phân loại và gắn thẻ, Thống kê và báo cáo, Tương tác người dung

2.4.3.2.Top Down Dashboard

Top Down Dashboard là một công cụ quản lý hiển thị tổng quan và chi tiết về tình hình và hoạt động của dự án, tổ chức hoặc doanh nghiệp. Nó cung cấp cái nhìn từ mức cao nhất xuống các chi tiết cụ thể, giúp người sử dụng nhận ra các vấn đề và cơ hội và theo dõi tiến độ theo góc nhìn toàn diện.

2.4.3.3.Bottom Up Dashboard

Bottom Up Dashboard là một công cụ quản lý hiển thị thông tin từ mức chi tiết lên đến tổng quan, thường được sử dụng để theo dõi và đánh giá tiến độ dự án, quản lý tài chính và các khía cạnh khác của hoạt động. Nó tập trung vào các chi tiết cụ thể và từ đó xây dựng cái nhìn tổng quan về tình hình và hiệu suất.

2.4.3.4. KPI Dashboard

KPI Dashboard (Key Performance Indicator Dashboard) là một công cụ quản lý hiển thị các chỉ số quan trọng (KPIs) một cách trực quan và dễ hiểu. Mục tiêu chính của KPI Dashboard là cung cấp cái nhìn tổng quan về hiệu suất và tiến độ của một dự án, tổ chức hoặc doanh nghiệp dựa trên các chỉ số quan trọng đã được xác định. Nó giúp người sử dụng theo dõi và đánh giá mức độ đạt được của các mục tiêu và kết quả quan trọng.

2.4.3.5. One Big Chart Dashboard

One Big Chart Dashboard là một loại bảng điều khiển dựa trên cơ sở của một biểu đồ lớn, thường là biểu đồ tương tác hoặc đồ thị có kích thước lớn và chi tiết. Mục tiêu chính của One Big Chart Dashboard là trình bày thông tin quan trọng và phức tạp trong một cái nhìn duy nhất bằng cách tập trung vào một biểu đồ chính.

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:

* Khám phá thông tin ẩn: Khai phá dữ liệu giúp tiết lộ thông tin, mẫu và tri thức ẩn chưa biết từ dữ liệu, giúp tăng cường hiểu biết về dự đoán, xu hướng và mối quan hệ.
* Dự đoán và dự báo: Khai phá dữ liệu có thể dự đoán xu hướng tương lai và kết quả dựa trên dữ liệu lịch sử, giúp trong quá trình ra quyết định kinh doanh và lập kế hoạch.
* Phát hiện thông tin giá trị: Việc tìm kiếm thông tin quan trọng giúp tìm ra những dấu hiệu, sự kiện quan trọng hoặc cơ hội trong dữ liệu.
* Tối ưu hóa quy trình: Khai phá dữ liệu có thể giúp cải thiện quy trình hoạt động và tối ưu hóa hiệu suất, chẳng hạn như tối ưu hóa cách tổ chức sản xuất hoặc cung ứng.
* Hiểu biết về khách hàng: Khai phá dữ liệu giúp xây dựng hồ sơ khách hàng chi tiết, từ đó cung cấp thông tin quan trọng để tùy chỉnh sản phẩm và dịch vụ.
* Phân tích thị trường: Khai phá dữ liệu hỗ trợ trong việc phân tích thị trường, xác định xu hướng mua sắm và ưa thích của khách hàng.
* Phát triển chiến lược kinh doanh: Thông qua việc phân tích dữ liệu, khai phá dữ liệu giúp hỗ trợ quá trình xác định và thực hiện chiến lược kinh doanh hiệu quả.
* Phân loại và nhóm dữ liệu: Khai phá dữ liệu giúp phân loại dữ liệu vào các nhóm tương tự nhau, từ đó giúp dễ dàng quản lý và sử dụng.

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 dữ liệu thành các nhóm tương tự nhau dựa trên các đặc điểm chung, giúp hiểu về cấu trúc dữ liệu và tương tác giữa các nhóm.
* **Phân loại (Classification):** Xây dựng mô hình để phân loại dữ liệu mới vào các lớp hoặc nhãn đã biết từ dữ liệu huấn luyện.
* **Hồi quy (Regression):** Dự đoán giá trị của một biến phụ thuộc dựa trên các biến độc lập, giúp hiểu mối quan hệ giữa các biến.
* **Khai thác liên kết (Association Rule Mining):** Phát hiện mẫu liên kết giữa các mục trong tập dữ liệu, giúp hiểu về mối tương quan và quy luật kết hợp.
* **Phát hiện bất thường (Anomaly Detection):** Tìm kiếm các quan sát không bình thường hoặc bất thường trong dữ liệu, hữu ích trong việc phát hiện gian lận hoặc sự cố.
* **Phân tích chuỗi thời gian (Time Series Analysis):** Phân tích xu hướng, mô hình và dự đoán giá trị trong chuỗi thời gian, thường được sử dụng trong dữ liệu thời gian thực.
* **Phân tích đồ thị (Graph Analysis):** Xác định mối quan hệ và cấu trúc của dữ liệu trong đồ thị, thường được áp dụng trong mạng xã hội hoặc mạng máy tính.
* **Phân tích văn bản (Text Mining):** Khai phá thông tin từ dữ liệu văn bản, bao gồm phân loại văn bản, phát hiện ý định, và phân tích tư duy của người dùng.
* **Phân tích tương tự (Similarity Analysis):** Đo lường sự tương tự giữa các mục dựa trên các đặc trưng hoặc thuộc tính chung.
* **Khai phá dữ liệu trực quan (Visual Data Mining):** Sử dụng biểu đồ và hình ảnh để hiển thị thông tin phức tạp trong dữ liệu.

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.

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

* Chọn thuộc tính quan trọng: Lựa chọn thuộc tính tốt nhất để kiểm tra tại mỗi nút dựa trên mức độ phân loại hoặc phân chia dữ liệu.
* Tách dữ liệu: Tách dữ liệu thành các tập con dựa trên các giá trị thuộc tính được chọn.
* Xây dựng nút quyết định: Tạo nút quyết định và liên kết chúng với các tập con dữ liệu tương ứng.
* Lặp lại các bước trên: Lặp lại quá trình trên cho các nút con cho đến khi một điều kiện dừng được đáp ứng.
* Cắt tỉa (Pruning): Loại bỏ những phần cây không quan trọng hoặc có thể gây quá mức phức tạp cho mô hình.
* 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): Chọn một số nguyên dương K, đại diện cho số lượng điểm láng giềng mà bạn muốn sử dụng để dự đoán.
* Tính khoảng cách: Tính khoảng cách giữa điểm dữ liệu mới và tất cả các điểm trong tập dữ liệu huấn luyện, sử dụng một phép đo khoảng cách như khoảng cách Euclidean.
* Chọn K láng giềng gần nhất: Chọn K điểm có khoảng cách nhỏ nhất với điểm dữ liệu mới.
* Phân loại hoặc dự đoán: Dự đoán lớp của điểm dữ liệu mới bằng cách chọn lớp phổ biến nhất trong K láng giềng (trong trường hợp phân loại) hoặc tính trung bình giá trị của K láng giềng (trong trường hợp dự đoán giá trị).

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 nhiễu dữ liệu và yêu cầu một lượng lớn dữ liệu huấn luyện để đạt được hiệu suất tốt.

**2.5.4.3. Random Forest:**

Random 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:

* **Phân loại và dự đoán:** Random Forest có thể được sử dụng cho cả bài toán phân loại (classification) và dự đoán (regression).
* Khả năng xử lý nhiễu: Do sử dụng nhiều cây quyết định, Random Forest có khả năng xử lý tốt dữ liệu nhiễu và thiếu sót.
* Tích hợp các cây quyết định: Bằng cách kết hợp dự đoán từ nhiều cây con, Random Forest cung cấp một dự đoán tổng quát mạnh mẽ.
* Tính hiệu suất: Random Forest thường có hiệu suất tốt trên nhiều loại dữ liệu và không yêu cầu nhiều điều chỉnh thủ công.
* Tránh overfitting: Bằng cách sử dụng các tập dữ liệu con và đặc trưng ngẫu nhiên, Random Forest giảm thiểu nguy cơ overfitting so với một cây quyết định đơn lẻ.

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

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.

Các khái niệm trong Confusion Matrix:

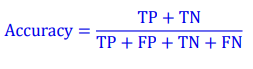
* **True Positive (TP):** Số lượng trường hợp mô hình dự đoán là Positive và thực tế cũng là Positive.
* **False Positive (FP):** Số lượng trường hợp mô hình dự đoán là Positive nhưng thực tế là Negative (nhầm lẫn).
* **True Negative (TN):** Số lượng trường hợp mô hình dự đoán là Negative và thực tế cũng là Negative.
* **False Negative (FN):** Số lượng trường hợp mô hình dự đoán là Negative nhưng thực tế là Positive (nhầm lẫn).

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).

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.

2.5.5.2. Các chỉ số đánh giá mô hình

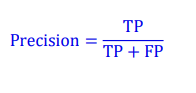
* Accuracy: Accuracy (độ chính xác) là một chỉ số đánh giá hiệu suất của mô hình phân loại trong học máy. Nó đo lường tỷ lệ phần trăm của số lượng dự đoán đúng (cả Positive và Negative) so với tổng số lượng dự đoán. Accuracy là một trong những chỉ số đơn giản và thường được sử dụng để đánh giá mô hình, nhưng cũng có thể bị ảnh hưởng bởi sự không cân bằng giữa các lớp hoặc nhãn



Độ chính xác càng cao thì mô hình càng tốt trong việc dự đoán. Tuy nhiên, độ chính xác không phản ánh rõ ràng khả năng dự đoán cho từng lớp cụ thể và có thể không thích hợp trong trường hợp dữ liệu không cân bằng, khi một lớp chiếm đa số trong tập dữ liệu. Trong các tình huống như vậy, các chỉ số khác như Precision, Recall và F1-score có thể cung cấp cái nhìn tổng quát hơn về hiệu suất của mô hình.

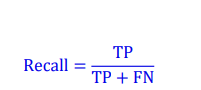
* **Precision**

Precision: Precision (độ chính xác dương tính) là một chỉ số quan trọng trong đánh giá hiệu suất của mô hình phân loại, đặc biệt là khi chúng ta quan tâm đến việc dự đoán lớp Positive (cũng gọi là dương tính). Precision đo lường tỷ lệ giữa số lượng dự đoán Positive đúng (True Positives) và tổng số lượng dự đoán Positive (bao gồm cả True Positives và False Positives).



* **Recall**

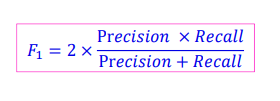
Recall đo lường tỷ lệ dự báo chính xác các trường hợp Positive trên toàn bộ các mẫu thuộc nhóm postive. Công thức như sau:



Để tính được recall thì chúng ta phải biết trước nhãn của dữ liệu. Do đó recall có thể được dùng để đánh giá trên tập có trước về kết quả đầu ra vì chúng ta đã biết trước nhãn. Trên tập test khi dữ liệu được coi như mới hoàn toàn và chưa biết nhãn thì chúng ta sẽ sử dụng precison.

* **F1-Score**

F1-Score (F1-score hoặc F1-measure) là một chỉ số kết hợp giữa Precision và Recall để đánh giá hiệu suất của mô hình phân loại trong học máy. Nó thường được sử dụng trong các trường hợp mô hình cần cân nhắc cả sự cân bằng giữa việc đạt được Precision cao và Recall cao.



F1-Score giúp cân nhắc giữa Precision và Recall. Nếu mô hình có cả Precision và Recall cao thì F1-Score sẽ cao. Nếu một trong hai chỉ số này thấp, F1-Score sẽ giảm.

F1-Score thường được ưa chuộng trong các tình huống mà chúng ta cần đảm bảo cân bằng giữa việc đạt được Precision cao và Recall cao.

2.5.5.3. AUC

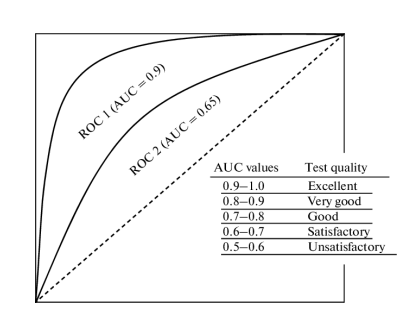
AUC (Area Under the Curve) là một chỉ số thường được sử dụng trong đánh giá hiệu suất của các mô hình phân loại trong học máy. Nó đo lường diện tích dưới đường cong ROC (Receiver Operating Characteristic), là biểu đồ biểu thị mối quan hệ giữa tỷ lệ True Positive Rate (Recall) và tỷ lệ False Positive Rate (1 - Specificity) khi ngưỡng dự đoán thay đổi.

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.

Một giá trị AUC cụ thể có thể diễn đạt như sau:

* AUC = 0.5: Mô hình không có khả năng phân loại tốt hơn một cách ngẫu nhiên.
* AUC > 0.5 và gần 1: Mô hình có khả năng phân loại tốt và tốt hơn càng gần 1.
* AUC = 1: Mô hình phân loại tuyệt vời, hoàn hảo.

AUC là một chỉ số hữu ích để so sánh hiệu suất của các mô hình phân loại và xác định xem mô hình nào tốt hơn trong việc phân loại các lớp khác nhau.



Hình 2.5.5.3 Minh họa AUC

1. **PHƯƠNG PHÁP THU THẬP VÀ XỬ LÝ DỮ LIỆU**

## 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 <https://www.kaggle.com/datasets/vanviethieuanh/vietnam-weather-data>

- Bộ dữ liệu gồm có: 10 cột và 181960 dòng

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) |

*Bảng 3.1.2 Cấu trúc và ý nghĩa bộ dữ liệu Weather*

## 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 một 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":

Thu thập dữ liệu: Bước đầu tiên là thu thập dữ liệu thời tiết từ các nguồn đáng tin cậy như trạm quan trắc, cơ quan khí tượng quốc gia, hoặc dịch vụ dự báo thời tiết trực tuyến. Dữ liệu này bao gồm thông tin về nhiệt độ, độ ẩm, áp suất không khí, tốc độ gió và các biến thời tiết khác.

Tiền xử lí dữ liệu: Dữ liệu thời tiết thường có độ biến đổi lớn và có thể bị nhiễu. Bước này bao gồm việc loại bỏ dữ liệu nhiễu, điền giá trị thiếu, và chuẩn hóa dữ liệu để đảm bảo tính nhất quán và chuẩn xác trong quá trình học.

Tạo các đặc trưng (features): Để mô hình học máy có thể hiểu dữ liệu thời tiết, cần tạo các đặc trưng thích hợp từ dữ liệu gốc. Điều này có thể bao gồm việc tạo các biến số thống kê như trung bình, độ biến đổi, hay cảm biến dữ liệu từ nhiều thời điểm trước đó để lưu trữ thông tin về xu hướng.

Chia dữ liệu thành tập huấn luyện và tập kiểm tra: Để đánh giá hiệu suất của mô hình, cần chia dữ liệu thành hai phần: tập huấn luyện để huấn luyện mô hình và tập kiểm tra để đánh giá khả năng dự đoán của mô hình trên dữ liệu chưa từng thấy.

Xây dựng mô hình học máy: Sử dụng các mô hình học máy như hồi quy tuyến tính, mạng nơ-ron, máy vector hỗ trợ (SVM), hay cây quyết định để huấn luyện và dự đoán dữ liệu thời tiết. Việc lựa chọn mô hình phụ thuộc vào đặc điểm của dữ liệu và mục tiêu của nghiên cứu.

Đánh giá và tinh chỉnh mô hình: Đánh giá hiệu suất của mô hình bằng các phép đo như độ chính xác, RMSE (Root Mean Squared Error), hay MAE (Mean Absolute Error). Dựa trên kết quả đánh giá, có thể thực hiện tinh chỉnh mô hình để cải thiện khả năng dự đoán.

Dự đoán và áp dụng: Khi mô hình đã được huấn luyện và tinh chỉnh, có thể sử dụng nó để dự đoán thời tiết trong tương lai. Thông tin dự đoán này có thể được ứng dụng trong nhiều lĩnh vực như giao thông, nông nghiệp, và quản lý tài nguyên.

Việc xử lí dữ liệu là bước quan trọng để đảm bảo dữ liệu thời tiết được chuẩn bị tốt để huấn luyện mô hình học máy và tạo ra các dự đoán chính xác về thời tiết.

# 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. Dưới đây là những yêu cầu quan trọng mà đề tài này đặt ra:

* Tăng cường độ chính xác của dự đoán: Mục tiêu chính của việc áp dụng mô hình học máy trong dự đoán thời tiết là tăng cường độ chính xác và độ tin cậy của dự đoán. Mô hình cần phải được xây dựng và tinh chỉnh để có khả năng dự đoán chính xác các yếu tố thời tiết như nhiệt độ, độ ẩm, áp suất không khí và tốc độ gió.
* Đối mặt với biến đổi khí hậu và thời tiết cực đoan: Đề tài đặt ra yêu cầu về khả năng dự đoán các biến đổi khí hậu và thời tiết cực đoan. Mô hình cần phải đáp ứng được trong các tình huống thời tiết đặc biệt như cơn bão, lũ lụt, nhiệt độ cực đoan và biến đổi khí hậu.
* Xử lí dữ liệu không chắc chắn và nhiễu: Dữ liệu thời tiết thường có độ biến đổi lớn và có thể bị nhiễu. Mô hình cần phải xử lí dữ liệu không chắc chắn và nhiễu để tạo ra các dự đoán ổn định và chính xác.
* Ứng dụng thực tế và lợi ích đa dạng: Yêu cầu của đề tài là phải áp dụng mô hình học máy vào các lĩnh vực thực tế như giao thông, nông nghiệp, năng lượng và du lịch. Mô hình cần phải mang lại lợi ích đa dạng và cụ thể cho các lĩnh vực này.
* Độ tin cậy và khả năng ứng phó: Mô hình cần đảm bảo rằng các dự đoán thời tiết không chỉ chính xác mà còn có khả năng ứng phó trong trường hợp các biến đổi đột ngột. Điều này đặt ra yêu cầu về tính tin cậy và khả năng ứng phó của mô hình.
* Hiểu biết sâu hơn về dữ liệu thời tiết: Đề tài yêu cầu phải có sự hiểu biết sâu hơn về dữ liệu thời tiết, từ các biến thời tiết cơ bản đến các yếu tố phức tạp như biến đổi khí hậu và tương tác giữa các yếu tố.

Việc áp dụng mô hình học máy trong phân tích và dự đoán dữ liệu thời tiết đặt ra nhiều yêu cầu quan trọng nhằm nâng cao khả năng dự báo thời tiết và ứng dụng thực tế trong nhiều lĩnh vực quan trọng.

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



Tiếp theo , khai báo các thư viên của Python để sử dụng.Các thư viện có chức năng như:

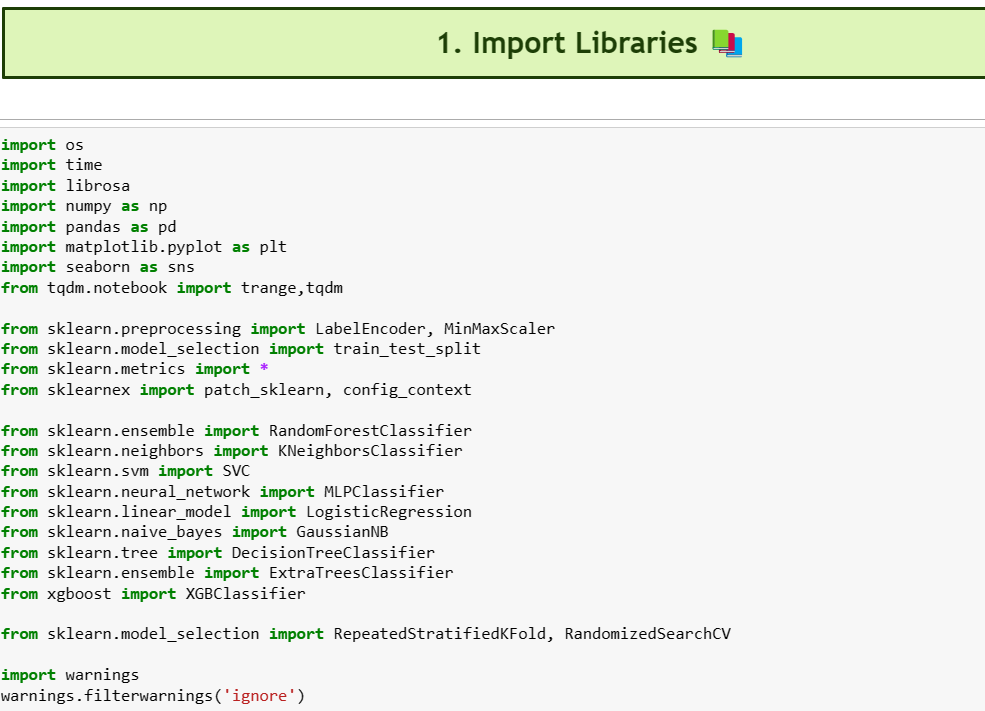
NumPy: cung cấp các phương pháp và công cụ tính toán số học và thống kê cho việc phân tích dữ liệu.

Pandas: thư viên hỗ trợ xử lý dữ liệu, giúp thực hiện các tác vụ như tải dữ liệu, chuyển đổi dữ liệu, xử lý dữ liệu thiếu,…

Matplotlib: thư viện đồ họa để vẽ các biểu đồ, biểu đồ đường, biểu đồ cột,..

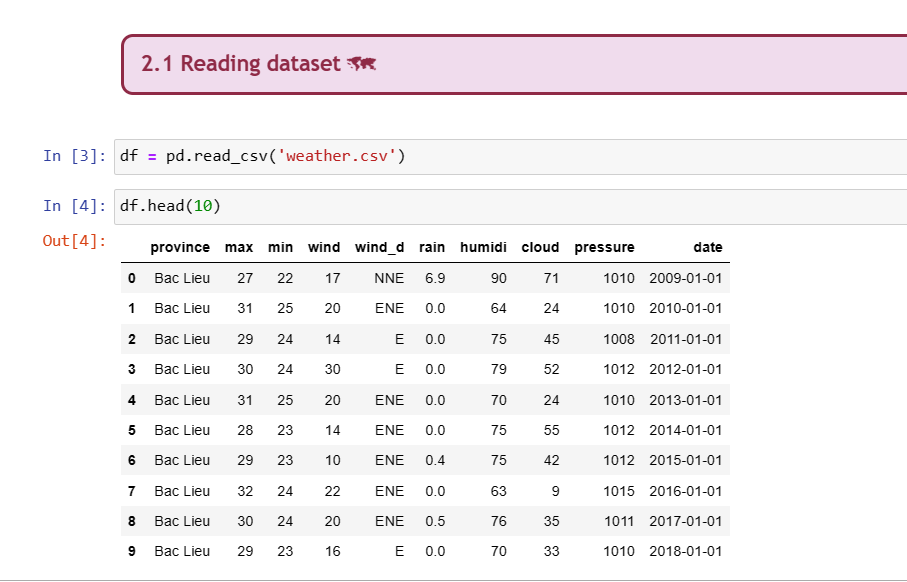
Seaborn: thư viện đồ họa dành cho việc phân tích dữ liệu, đặc biệt là trực quan hóa dữ liệu

Scikit-learn: thư viện hỗ trợ các thuật toán học máy cho các tác vụ như phân loại, dự báo và gom nhóm



*Hình 4.2 Khai báo các thư viên Python sử dụng*

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

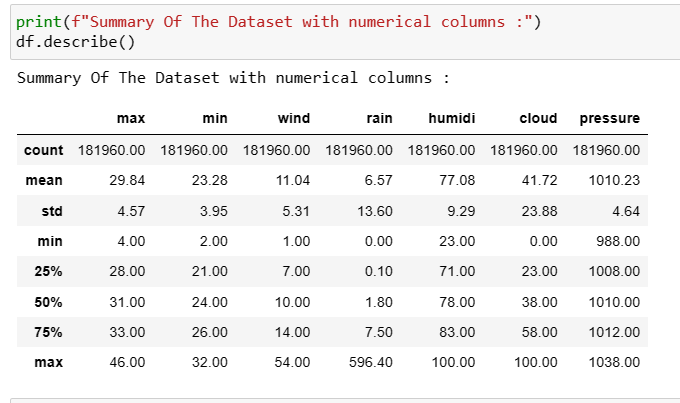


*Hình 4.3 Dữ liệu sau khi tải lên notebook*



*Hình 4.4 Xem dữ liệu có trong các côt*

File dữ liệu được sử dụng đã được xử lý, nhưng để tránh sai sót, cần thực hiện kiểm tra lại dữ liệu. Việc kiểm tra dữ liệu gồm: xem thông tin bảng dữ liệu, kiểm tra các giá trị trung bình, độ lệch chuẩn, giá trị nhỏ nhất, giá trị lớn nhất…

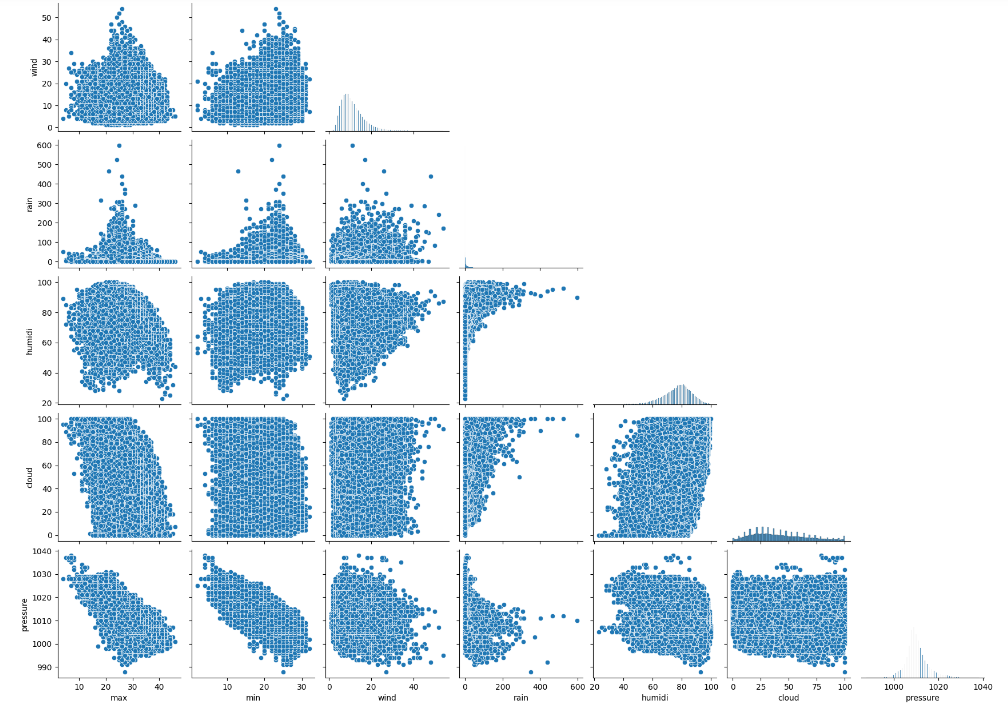


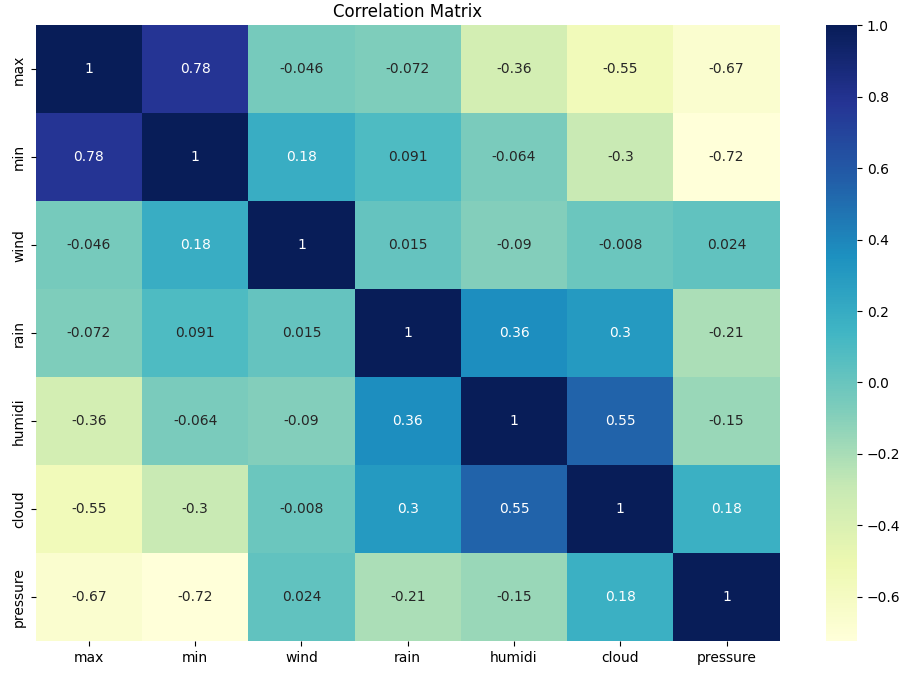
*Hình 4.5 Thống kê các thông số của dữ liệu*

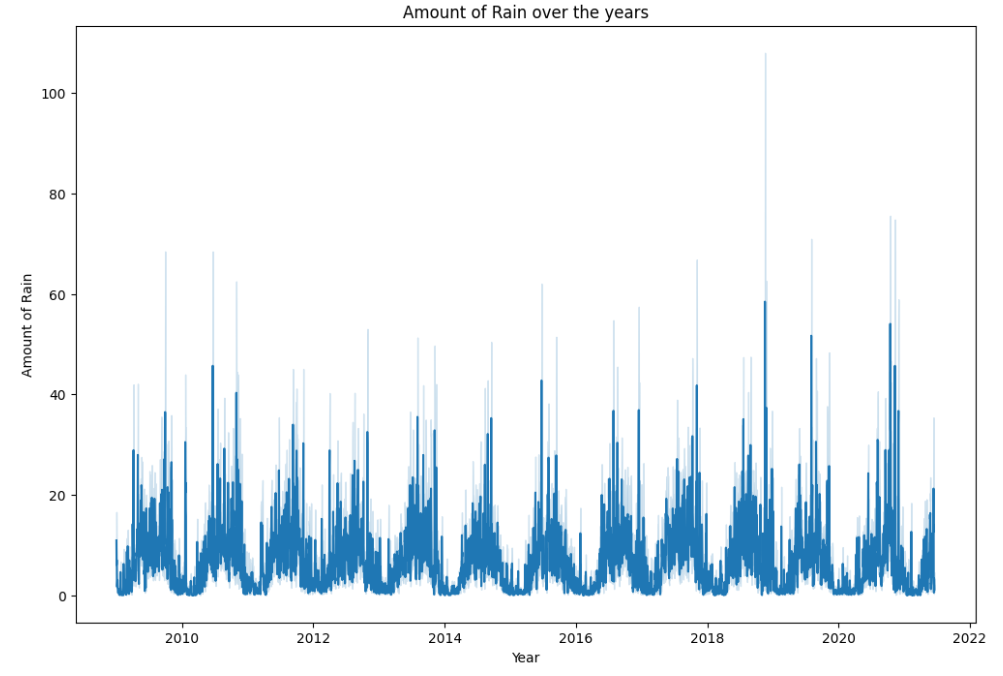
4.3. Xây dựng mô hình

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.

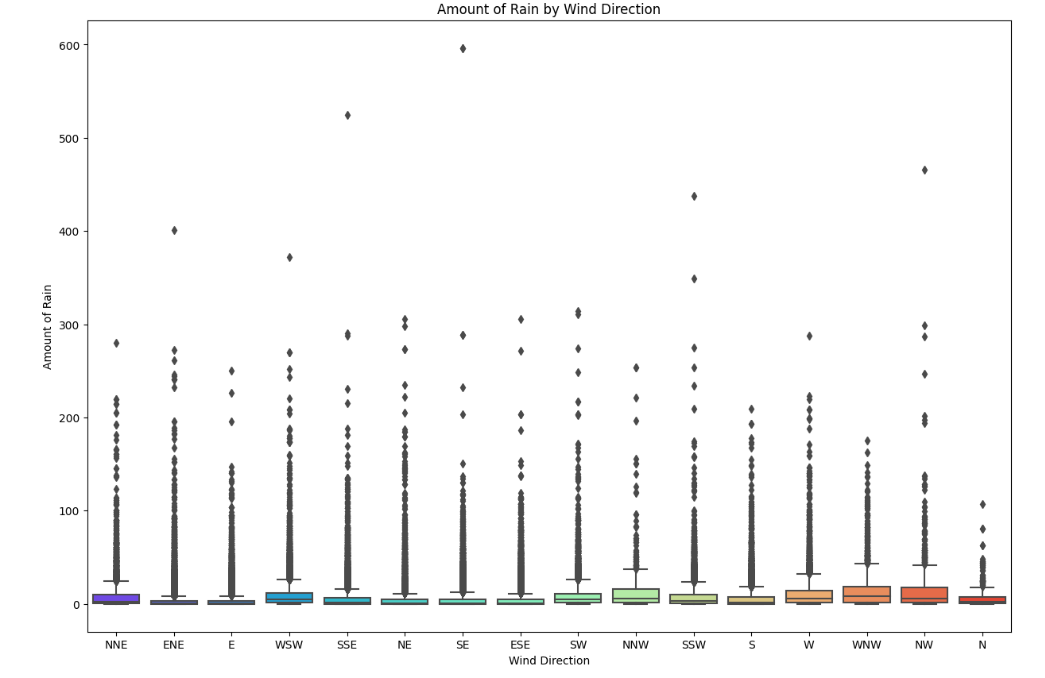
Tạo ma trận tương quan giữa các biến trong bộ dữ liệu để đánh giá mối quan hệ tuyến tính giữa các biến đó. Thông qua ma trận tương quan, có thể xác định được những biến nào có mối quan hệ mạnh và cần được lựa chọn để xây dựng mô hình. Các giá trị trên cả trục dọc và trục ngang thể hiện đặc trưng của bộ dữ liệu. Khi trực quan hóa ma trận tương quan, phương pháp sử dụng màu sắc để biểu thị mức độ tương quan. Màu tối hơn được sử dụng để chỉ mối tương quan dương cao, tức là khi hai đặc trưng tăng cùng nhau. Màu sáng hơn được sử dụng để chỉ mối tương quan âm cao hơn, tức là khi một đặc trưng tăng thì đặc trưng còn lại giảm



*Hình 4.6 Minh họa Correlation Matrix*



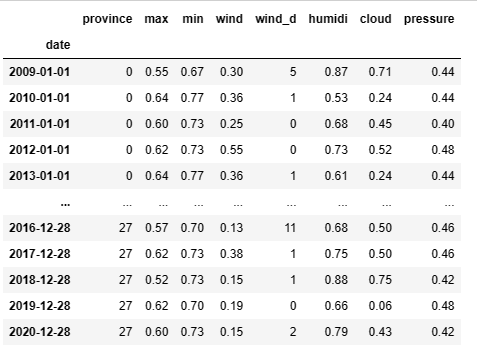
*Hình 4.7 Lượng mưa qua các năm*



*Hình 4.8 Lượng mưa theo hướng gió*

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.

Trước khi xây dựng mô hình, chúng ta cần mã hóa dữ liệu. Trong dự án này, em sử dụng hai kĩ thuật Min-max Scale cho dữ liệu kiểu numeric và Label Encode cho dữ liệu kiểu object. Dữ liệu sau khi mã hóa được biểu diễn như bên dưới:



Hình 4.9 Chuyển đổi từ object sáng numeric

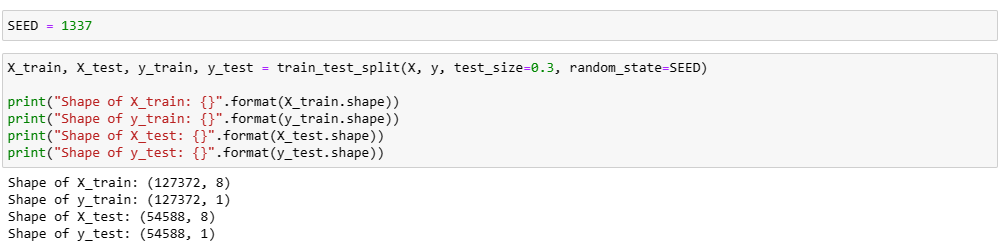
Chúng ta thấy column “province” và “wind\_d” với kiểu dữ liệu object tương ứng với tên Tỉnh thành và hướng gió được chuyển đổi từ kiểu object sang numeric. Và các column còn lại tương ứng với kiểu số được mã hóa dựa trên kĩ thuật Min-Max Scale chuyển đổi dữ liệu kiểu số thành các giá trị trong khoảng từ 0 đến 1.

Sau khi Normalize dữ liệu, chúng ta cùng đến với bước xác định biến dự đoán và phân chia tập train, test. Sau khi phân tích và nghiên cứu, đã chọn được biến cần dự đoán là “lượng mưa (rain)” với mô hình dự đoán Classification.



*Hình 4.10 chạy vòng lặp for*

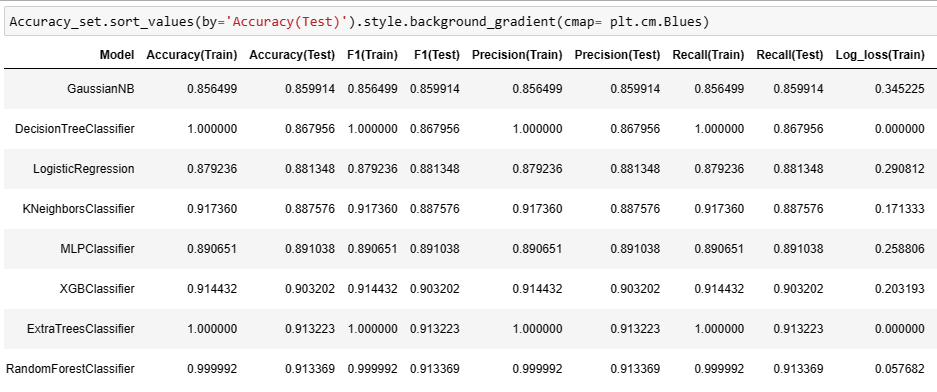
Đoạn code trên chuyển đổi những giá trị lớn hơn 0 của biến “rain” thành 1 và các giá trị bằng 0 thành 0. Có nghĩa là bài toán trở thành bài toán phân loại thời tiết mưa hoặc không mưa. Sau đó chúng ta chia tập train test:



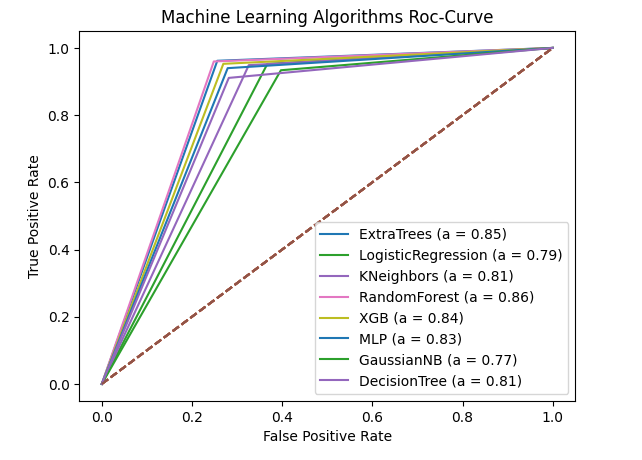
*Hình 4.11 Tập huấn luyện train và tập kiểm tra test*

Tập huấn luyện (train) và tập kiểm tra (test) với tỷ lệ dữ liệu kiểm tra là 30% (test\_size = 0.3). Điều này có nghĩa là chúng ta lấy 70% dữ liệu để huấn luyện mô hình và 30% dữ liệu được sử dụng để đánh giá và kiểm tra hiệu suất của mô hình. Gía trị ngẫu nhiên random\_state để đảm bảo quá trình chia dữ liệu được diễn ra một cách ngẫu nhiên.

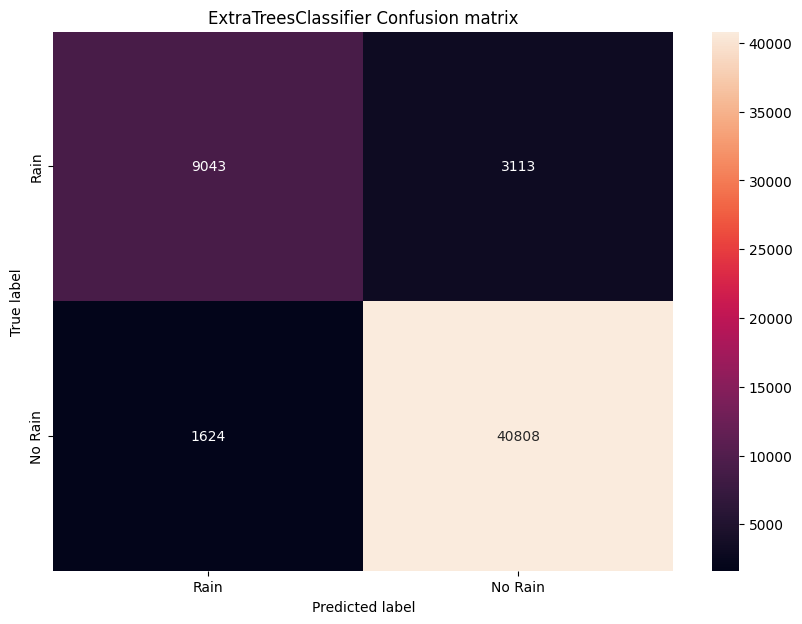
Sau khi thực hiện quá trình train, test, mô hình đạt được những kết quả sau đây:



*Hình 4.12 Kết quả của quá trình chạy train và test*

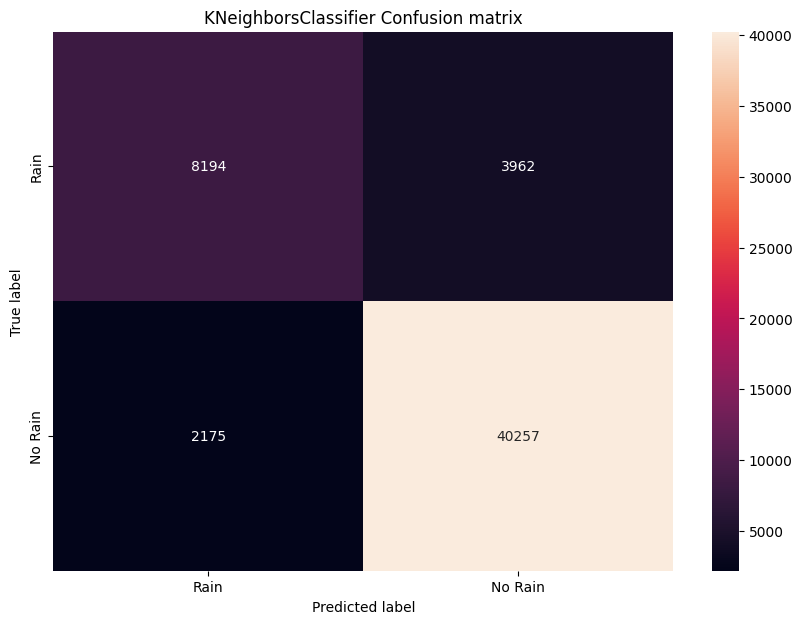


*Hình 4.13 Kết quả của đồ thị ROC*

 *Hình 4.14 Ma trận Confusion Matrix của mô hình ExtraTreesClassifier*

* True positive (TP): Số dự đoán đúng thời tiết sẽ mưa là 9043
* True negative (TN): Số dự đoán đúng thời tiết sẽ không mưa là 40808
* False positive (FP): Số dự đoán thời tiết sẽ không mưa nhưng thực tết là mưa là 3113
* False negative (FN): Số dự đoán thời tiết sẽ mưa nhưng thực tết là không mưa là

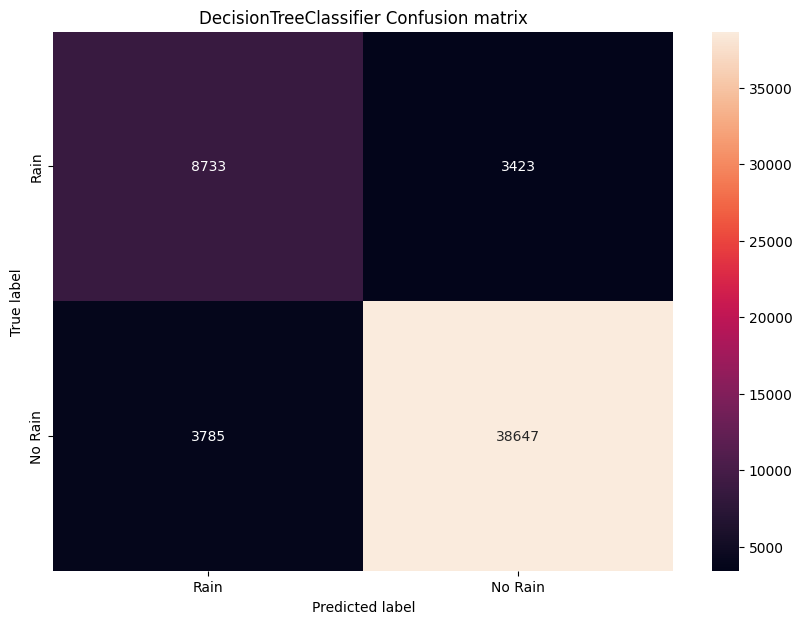
1624



*Hình 4.15 Ma trận Confusion Matrix của mô hình KNeighbors Classifier*

* True positive (TP): Số dự đoán đúng thời tiết sẽ mưa là 8194
* True negative (TN): Số dự đoán đúng thời tiết sẽ không mưa là 40257
* False positive (FP): Số dự đoán thời tiết sẽ không mưa nhưng thực tết là mưa là 3962
* False negative (FN): Số dự đoán thời tiết sẽ mưa nhưng thực tết là không mưa là

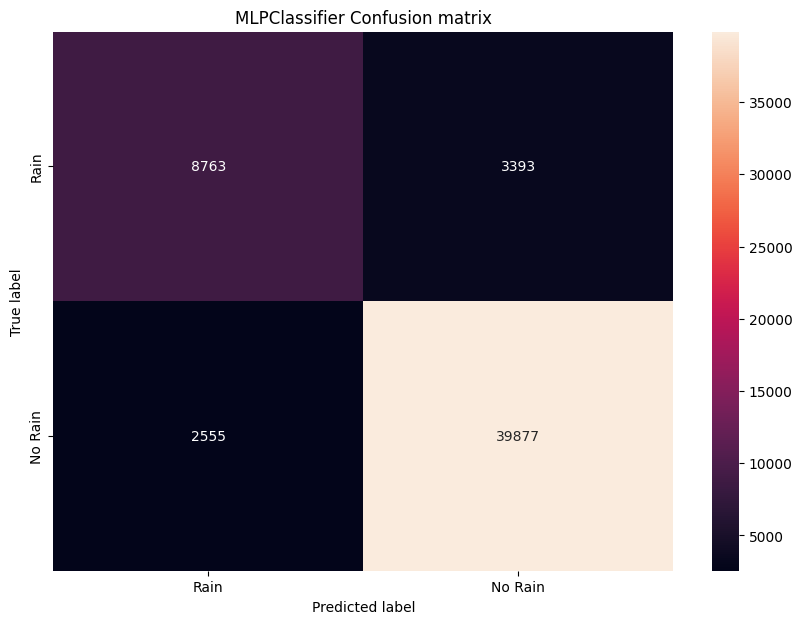
2175



*Hình 4.16 Ma trận Confusion Matrix của mô hình Decision TreeClassifier*

* True positive (TP): Số dự đoán đúng thời tiết sẽ mưa là 8733
* True negative (TN): Số dự đoán đúng thời tiết sẽ không mưa là 38647
* False positive (FP): Số dự đoán thời tiết sẽ không mưa nhưng thực tết là mưa là 3423
* False negative (FN): Số dự đoán thời tiết sẽ mưa nhưng thực tết là không mưa là

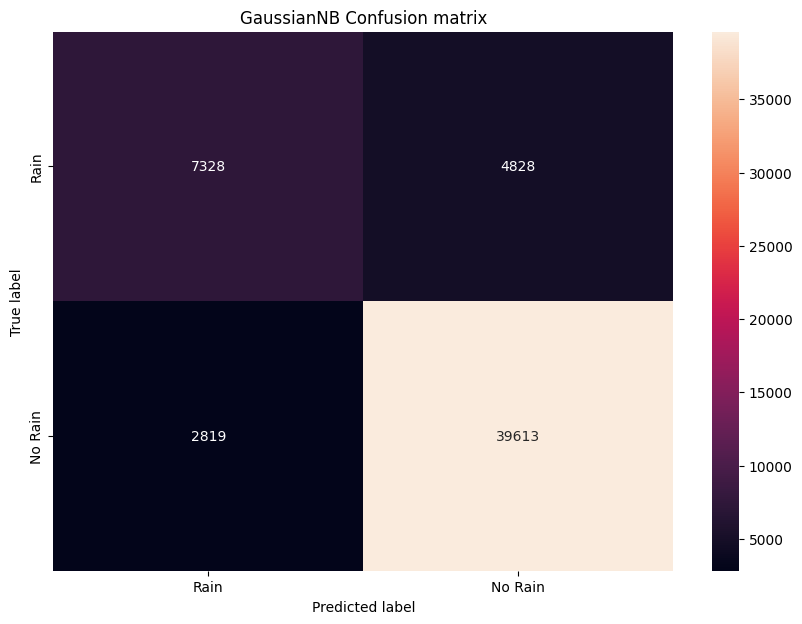
3785



*Hình 4.17 Ma trận Confusion Matrix của mô hình MLP Classifier*

* True positive (TP): Số dự đoán đúng thời tiết sẽ mưa là 8763
* True negative (TN): Số dự đoán đúng thời tiết sẽ không mưa là 39877
* False positive (FP): Số dự đoán thời tiết sẽ không mưa nhưng thực tết là mưa là 3393
* False negative (FN): Số dự đoán thời tiết sẽ mưa nhưng thực tết là không mưa là

2555

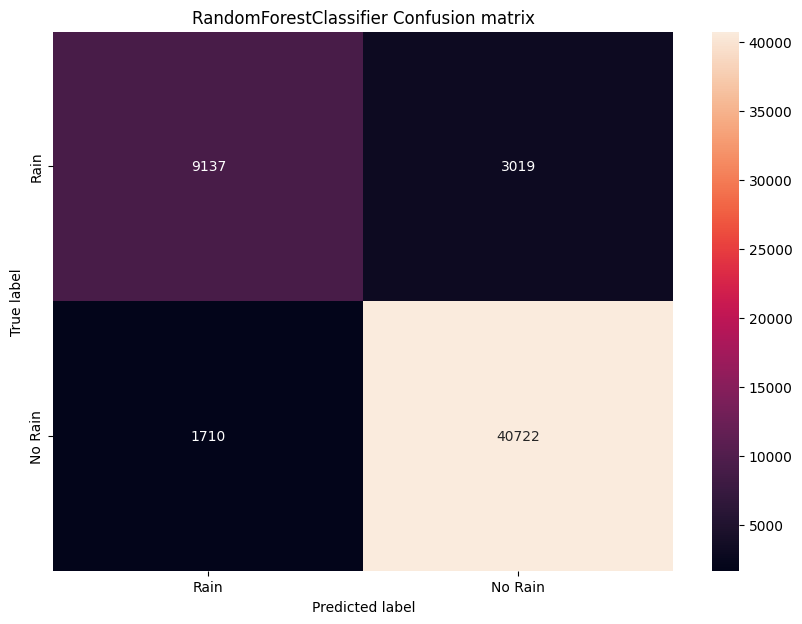


*Hình 4.18 Ma trận Confusion Matrix của mô hình GaussianNB Confusion*

* True positive (TP): Số dự đoán đúng thời tiết sẽ mưa là 7328
* True negative (TN): Số dự đoán đúng thời tiết sẽ không mưa là 39613
* False positive (FP): Số dự đoán thời tiết sẽ không mưa nhưng thực tết là mưa là 4828
* False negative (FN): Số dự đoán thời tiết sẽ mưa nhưng thực tết là không mưa là

2819

Sau khi so sánh và đánh giá chúng ta thấy rằng mô hình Random Forest cho ra kết quả tốt nhất với 91.34% ở tập test, giá trị AUC: 86% cùng với Confusion matrix sau:



*Hình 4.19 Ma trận Confusion Matrix của mô hình Random Forest 1Classifier*

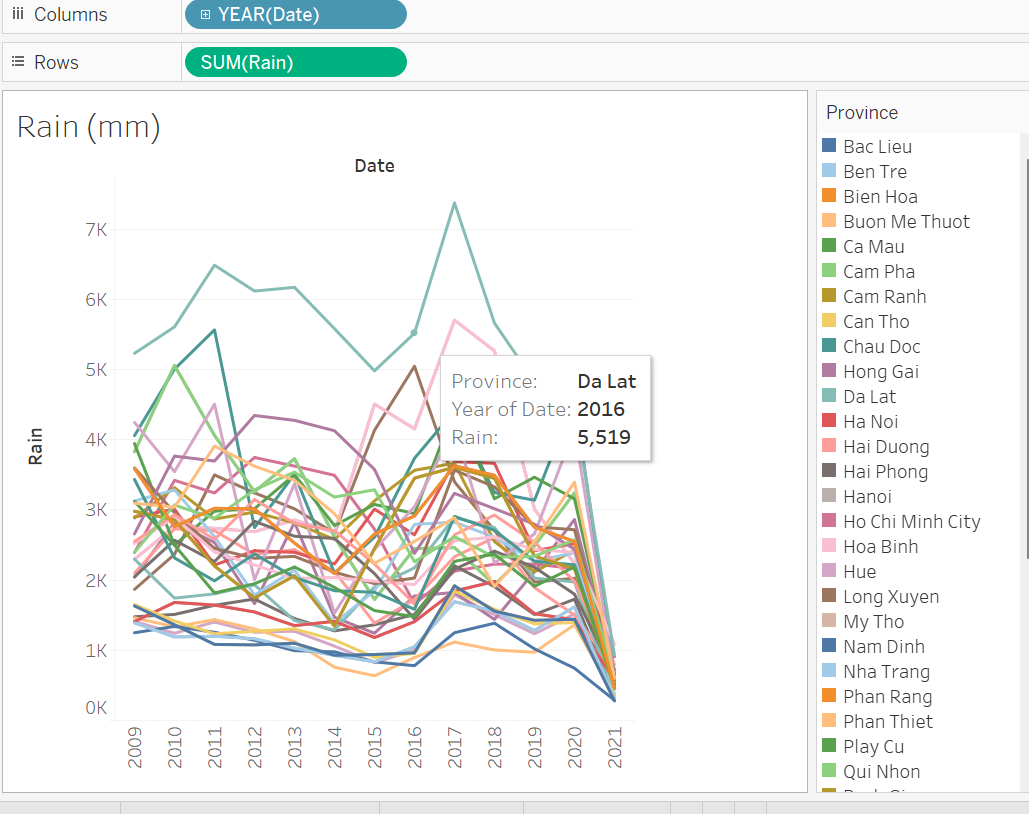
* True positive (TP): Số dự đoán đúng thời tiết sẽ mưa là 9137
* True negative (TN): Số dự đoán đúng thời tiết sẽ không mưa là 40722
* False positive (FP): Số dự đoán thời tiết sẽ không mưa nhưng thực tết là mưa là 3019
* False negative (FN): Số dự đoán thời tiết sẽ mưa nhưng thực tết là không mưa là

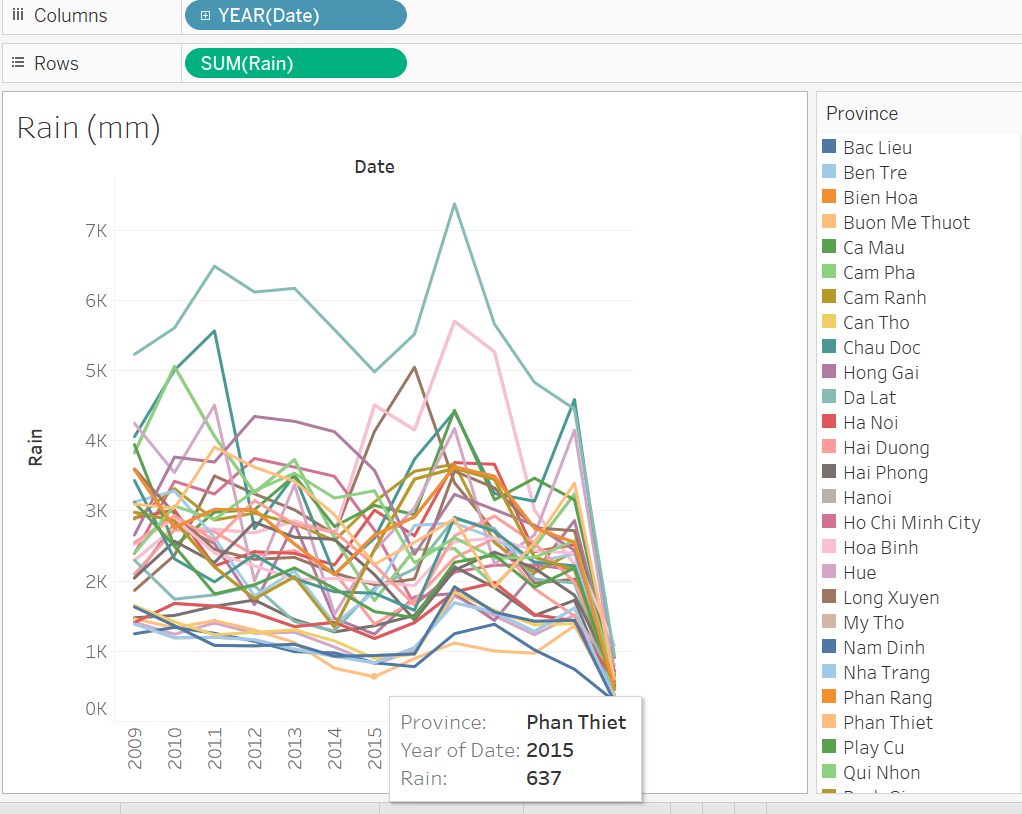
1710

* Từ bảng Confusion Matrix ta tính được các giá trị F1, Precision, Recall xấp xỉ 91.34%. Mô hình dự đoán rất tốt

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

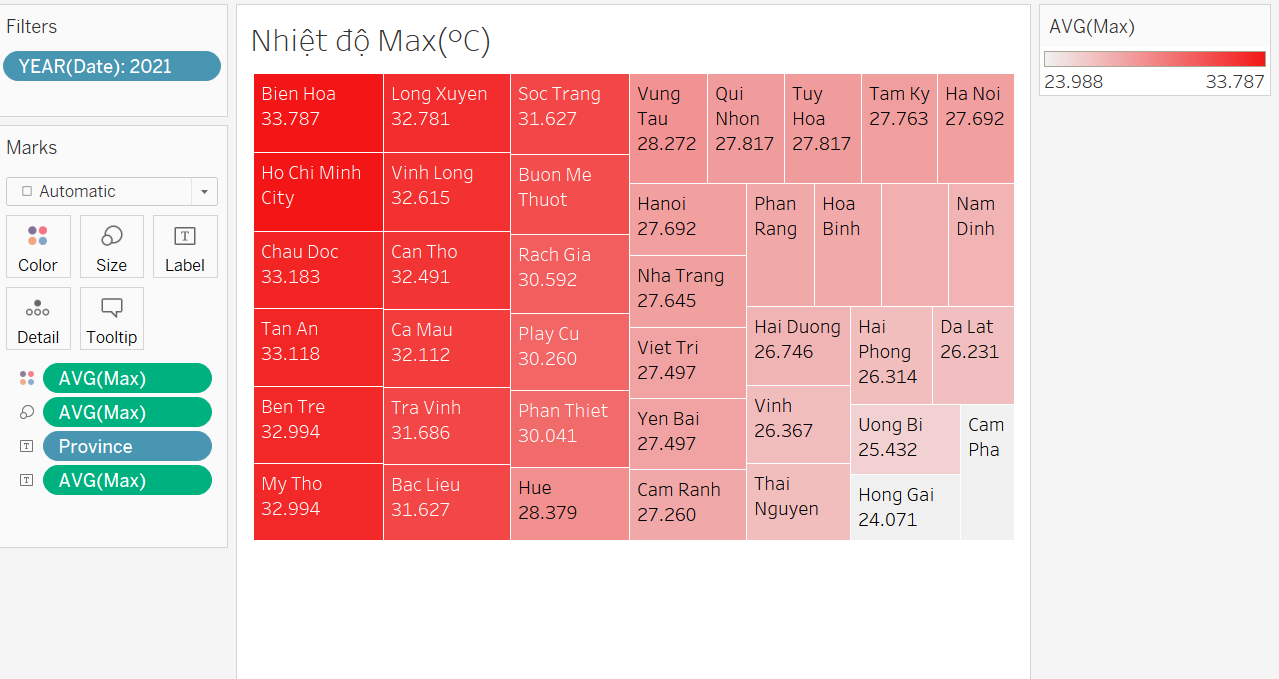
* Biểu đồ thể hiện tổng quan lượng mưa của các tỉnh





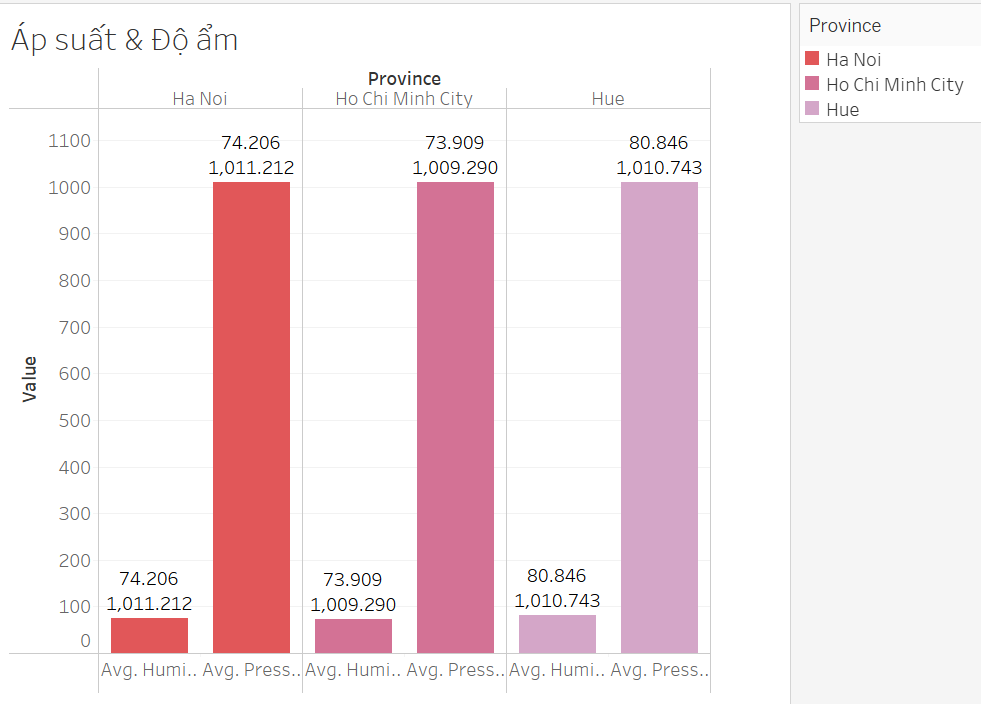
*Hình 4.4 Biểu đồ tổng lượng mưa ở các Tỉnh trong Nước Việt Nam*

* Nhận xét:
  + Biểu đồ thể hiện tổng lượng mưa của các tỉnh dễ nhìn tỉnh có lượng mưa nhiều nhất ở đây là Đà Lạt nơi có khí hậu vô cùng mát mẻ.
  + Lượng mưa ít nhất ở tỉnh Phan Thiết
  + Qua 2 biểu đồ trên ta biết được lượng mưa , và thời tiết của các tỉnh như thế nào từ đó có những biện pháp phòng chóng mưa và lũ lụt. Nếu lượng mưa nhiều quá thì mọi người ở các tỉnh này hay đi du lịch sẽ cần phải trang bị dụng cụ , đồ dung thích hợp để tránh những trường hợp xấu xảy ra.



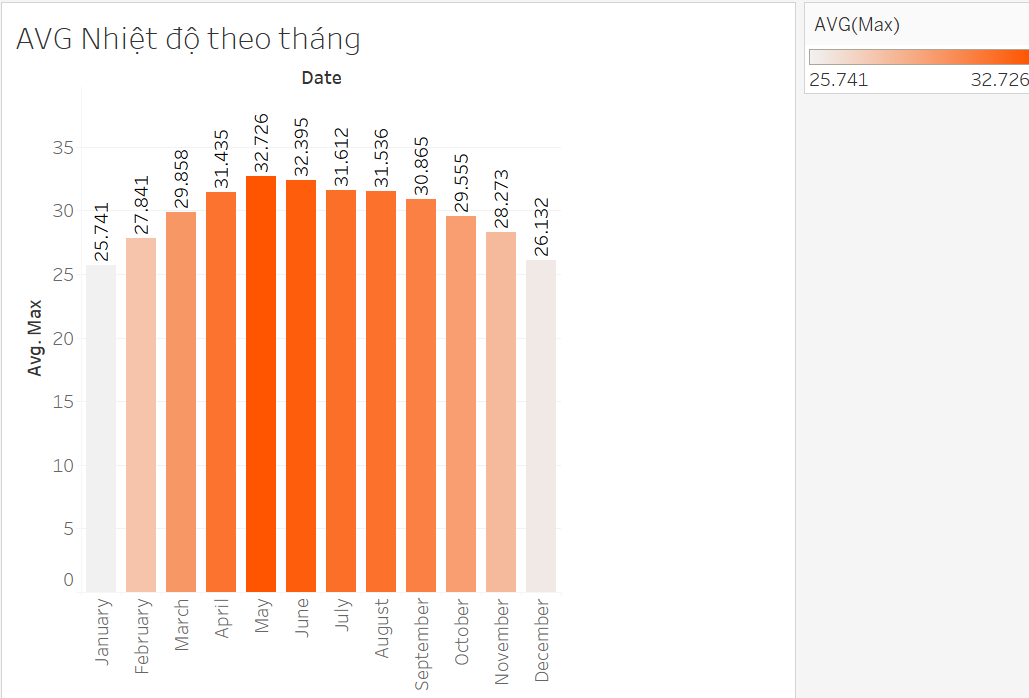
*Hình 4.5 Biểu đồ trung bình nhiệt độ cao nhất ở các tỉnh*

* Nhận xét: Ta thấy nơi có nhiệt độ trung bình cao nhất ở trên biểu đồ là tỉnh Biên Hòa với 33,787 °C và nơi có nhiệt độ thấp nhất là Cẩm Phả 23,288 °C
* Biện Pháp: Để giúp các tỉnh tránh khỏi tình trạng thời tiết nhiệt độ cao và tạo ra môi trường sống an toàn hơn trong những ngày nhiệt đới, có thể thực hiện một số biện pháp sau:
* Thông tin cảnh báo và tư vấn:Các cơ quan khí tượng cần thường xuyên cung cấp thông tin cảnh báo về thời tiết nhiệt đới cao cho các tỉnh. Các cơ quan này nên cung cấp tư vấn về cách ứng phó với nhiệt độ cao như cách bảo vệ sức khỏe, giữ đủ nước, và hạn chế hoạt động ngoài trời trong khoảng thời gian nhiệt độ cao nhất.
* Hạ nhiệt độ trong nhà:Khuyến khích người dân sử dụng các biện pháp như máy lạnh, quạt, và việc tạo thông gió trong nhà để giảm nhiệt độ bên trong. Cơ quan chính trị địa phương có thể hỗ trợ người dân có điều kiện hơn để sử dụng các thiết bị làm mát trong những ngày nhiệt độ cao.
* Tổ chức các điểm mát mẻ: Thiết lập các điểm mát mẻ như phòng khách công cộng, thư viện, trung tâm thể dục hoặc trung tâm thương mại để cung cấp nơi trú ẩn và thoải mái cho người dân trong thời gian nhiệt độ cao.
* Chăm sóc đặc biệt cho nhóm nguy cơ: Tập trung vào việc chăm sóc đặc biệt cho người già, trẻ em, và những người có vấn đề sức khỏe nghiêm trọng. Đảm bảo rằng họ được cung cấp nước đầy đủ và không phải tiếp xúc trực tiếp với nhiệt độ cao ngoài trời.
* Giám sát và phân tích dữ liệu thời tiết:Các cơ quan khí tượng cần tiếp tục giám sát và phân tích dữ liệu thời tiết để cung cấp thông tin chính xác và kịp thời về nhiệt độ dự kiến. Điều này sẽ giúp cơ quan chính trị và dân cư chuẩn bị kế hoạch phản ứng hiệu quả.
* Tạo hạt nhân xanh và cải thiện môi trường:Thiết lập các khu vườn, công viên, và cố gắng cải thiện môi trường xung quanh bằng cách trồng cây cối và tạo bóng mát. Các hạt nhân xanh này có thể giúp làm giảm nhiệt độ toàn cục và tạo ra không gian thoải mái hơn cho cư dân.
* Giáo dục về sức khỏe và an toàn:Tổ chức các chương trình giáo dục và tăng cường nhận thức về cách duy trì sức khỏe và an toàn trong thời tiết nhiệt đới. Các biện pháp bảo vệ da, đồng hồ nước, và lựa chọn trang phục phù hợp có thể được tập trung trong các hoạt động giáo dục này.



*Hình 4.6 Biểu đồ Áp suất và độ ẩm của các 3 Tỉnh*

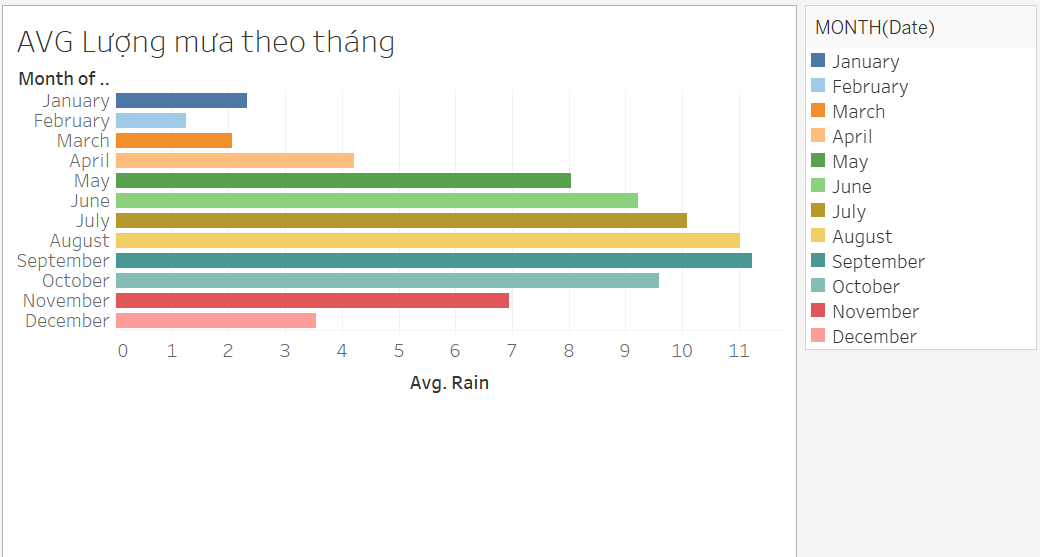
* Nhận xét: Ta có độ ẩm của 3 Tỉnh Hà Nội, Hồ Chí Minh , Huế lần lượt là : 74,206; 73,909 ; 80,864 và áp xuất là 1011212; 1009290; 1010743



*Hình 4.7 Biểu đồ nhiệt độ max theo tháng*

* Nhận xét: Ba tháng có nhiệt độ cao nhất ở đây là tháng 4,5,6 với nhiệt độ trung bình cao nhất lên đến 32,726 °C .
* Biện Pháp: Để giúp ứng phó với các tháng có nhiệt độ cao, có thể áp dụng một số biện pháp sau đây:
* Điều chỉnh hoạt động hàng ngày: Tránh thực hiện các hoạt động vật lý mạnh vào thời gian nhiệt độ cao nhất trong ngày. Chuyển các hoạt động ra ngoài buổi sáng sớm hoặc buổi tối khi nhiệt độ thấp hơn.
* Duy trì cân bằng nước: Uống đủ nước để duy trì cân bằng nước cơ thể trong thời tiết nhiệt đới. Tránh thức uống có chứa cafein và cồn, vì chúng có thể gây mất nước.
* Sử dụng quần áo mát mẻ: Mặc quần áo thoải mái, mỏng và mát mẻ để giúp cơ thể thoát nhiệt dễ dàng hơn.
* Bảo vệ da khỏi tác động của nhiệt độ cao: Sử dụng kem chống nắng và đội nón để bảo vệ da khỏi tác động của tia UV và nhiệt độ cao.
* Tạo bóng mát và thông gió: Sử dụng rèm cửa, màn cửa hoặc cửa sổ để che nắng và tạo bóng mát cho ngôi nhà. Mở cửa sổ để tạo thông gió và làm ngôi nhà mát mẻ hơn.
* Sử dụng máy lạnh và quạt: Sử dụng máy lạnh và quạt để làm giảm nhiệt độ bên trong nhà, tạo môi trường thoải mái cho gia đình.
* Chăm sóc đặc biệt cho người già và trẻ em: Đặc biệt chú trọng đến việc chăm sóc sức khỏe cho người già và trẻ em, bởi vì họ thường dễ bị ảnh hưởng bởi nhiệt độ cao.
* Giữ gìn vệ sinh: Thường xuyên tắm rửa và giữ vệ sinh cá nhân để tránh bị nhiễm trùng hoặc các vấn đề sức khỏe khác do nhiệt độ cao.
* Tìm nơi trú ẩn mát mẻ: Khi cần thiết, tìm nơi trú ẩn mát mẻ như thư viện, trung tâm thể dục hoặc cửa hàng để tránh tác động của nhiệt độ cao.
* Chế độ ăn uống hợp lý: Bổ sung thêm các loại thực phẩm có chứa nước, như rau cải và trái cây tươi, để giữ cho cơ thể luôn đủ nước và dưỡng chất trong thời tiết nhiệt đới.

Kết hợp những biện pháp trên sẽ giúp bạn và gia đình ứng phó tốt hơn với tháng có nhiệt độ cao, bảo vệ sức khỏe và tạo môi trường sống thoải mái hơn.



*Hình 4.8 Biểu đồ Lượng mưa theo tháng*

* Nhận xét: Ba tháng có lượng mưa cao nhất là 8,9, 10 với lượng mưa trung bình cao nhất lên đến 11,234 mm
* Biện Pháp: Để ứng phó với các tháng có lượng mưa cao và tăng cường khả năng quản lý dòng chảy nước, có thể áp dụng một số biện pháp sau đây:
* Xây dựng hệ thống thoát nước hiệu quả: Đảm bảo hệ thống thoát nước trong khu vực làm việc tốt, bao gồm các cống thoát nước, cống rãnh và hố chứa nước. Thực hiện công tác bảo dưỡng định kỳ để tránh tắc nghẽn và ngập úng.
* Rừng ngập mặn và cây xanh: Xây dựng và bảo vệ các khu vực rừng ngập mặn, cảnh quan cây xanh để hấp thụ lượng nước dư thừa và giảm thiểu tác động của lũ lụt.
* Xây dựng hệ thống chứa nước: Xây dựng các hồ chứa nước dự trữ để giữ lại lượng nước lớn trong thời gian mưa lớn. Các hồ này có thể giúp kiểm soát dòng chảy nước ra các con sông và hạn chế nguy cơ lũ lụt.
* Hệ thống dẫn nước: Xây dựng hệ thống đường ống dẫn nước từ các vùng có lượng mưa lớn đến các vùng cần nước để tận dụng nguồn nước dư thừa.
* Quản lý cảnh quan và xây dựng bền vững: Thực hiện quản lý cảnh quan thông minh, sử dụng các biện pháp bảo vệ đất đai như xây dựng bậc cảnh quan, cải tạo đất, và trồng cây trồng phủ.
* Hệ thống cảnh báo sớm: Phát triển hệ thống cảnh báo sớm về lượng mưa dự kiến, để cung cấp thông tin kịp thời cho cư dân và cơ quan chính trị để chuẩn bị ứng phó.
* Kế hoạch ứng phó dự lựa: Xây dựng kế hoạch ứng phó dự lựa cho các trường hợp lượng mưa cao đột ngột. Điều này bao gồm việc chuẩn bị các biện pháp sơ tán dân cư và đảm bảo nguồn cung cấp nước sạch.
* Hợp tác vùng địa phương: Hợp tác giữa các tổ chức chính trị, cơ quan khí tượng, chính phủ địa phương và cộng đồng dân cư để thực hiện các biện pháp ứng phó hiệu quả.
* Tầm nhìn bền vững: Thực hiện tầm nhìn bền vững trong quản lý tài nguyên nước và đất đai để giảm thiểu tác động của thay đổi khí hậu và tăng cường khả năng chịu đựng.
* Kết hợp những biện pháp trên sẽ giúp cải thiện khả năng quản lý lượng mưa cao, hạn chế tác động của lũ lụt và đảm bảo an toàn cho cư dân trong thời kỳ mưa lớn.

4.5. Đề xuất và ứng dụng

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

* Khuyến nghị sử dụng thông tin thời tiết trong đời sống hằng ngày để giúp bạn chuẩn bị tốt hơn cho mọi ngày và tối ưu hóa các hoạt động cá nhân. Dự báo thời tiết cung cấp thông tin quý báu về điều kiện nhiệt đới, mưa rào, hay nguy cơ bão, từ đó giúp bạn lựa chọn trang phục phù hợp, điều chỉnh lịch trình dự định, và tổ chức các hoạt động ngoài trời một cách thích hợp. Thông tin thời tiết còn giúp bạn bảo vệ sức khỏe và an toàn, bằng cách dự đoán các biến đổi không gian môi trường và thích nghi đúng cách.
* Ngoài ra, việc sử dụng thông tin thời tiết cũng mang lại lợi ích trong việc lập kế hoạch du lịch, tổ chức sự kiện, và quản lý thời gian cá nhân. Bạn có thể tránh những tình huống bất ngờ, giảm thiểu rủi ro cho các hoạt động ngoài trời, và tận dụng thời gian một cách hiệu quả hơn. Hơn nữa, việc tận dụng thông tin thời tiết giúp bạn có cái nhìn tổng quan về môi trường xung quanh, đồng thời thúc đẩy ý thức bảo vệ môi trường và áp dụng những biện pháp phù hợp trong cuộc sống hàng ngày.
* Việc sử dụng thông tin thời tiết trong đời sống là một cách thông minh để tận dụng thông tin hiện có để đảm bảo mọi hoạt động diễn ra một cách suôn sẻ, an toàn và hiệu quả.

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

Ứng dụng thông tin thời tiết vào các quyết định kinh doanh mang lại lợi ích vượt trội trong việc tối ưu hóa hoạt động và định hướng chiến lược. Bằng cách tích hợp dữ liệu thời tiết vào quá trình quản lý kinh doanh, doanh nghiệp có thể đạt được những kết quả sau:

* + - * Quản lý nguồn lực hiệu quả: Thông tin thời tiết giúp doanh nghiệp dự đoán nhu cầu thị trường và quản lý nguồn lực một cách hiệu quả. Ví dụ, ngành dịch vụ như nhà hàng có thể điều chỉnh thực đơn dựa trên dự báo thời tiết, từ đó giảm lãng phí và tối ưu hóa nguyên liệu.
      * Chiến lược tiếp thị đích đáng: Thông tin thời tiết ảnh hưởng đến hành vi mua sắm của khách hàng. Sử dụng thông tin này giúp doanh nghiệp xác định chiến lược tiếp thị tối ưu hóa dựa trên tình trạng thời tiết. Ví dụ, cửa hàng thời trang có thể quảng cáo sản phẩm phù hợp với thời tiết hiện tại.
* Quản lý chuỗi cung ứng: Thông tin thời tiết ảnh hưởng đến vận chuyển, lưu trữ và sản xuất. Doanh nghiệp có thể sử dụng dữ liệu thời tiết để dự đoán tình huống không lường trước, từ đó xây dựng kế hoạch ứng phó và backup cho chuỗi cung ứng.
* Nâng cao trải nghiệm khách hàng: Thông tin thời tiết có thể giúp cung cấp dịch vụ tốt hơn cho khách hàng. Ứng dụng di động có thể cung cấp dự báo thời tiết và gợi ý sản phẩm hoặc dịch vụ phù hợp dựa trên điều kiện thời tiết trong khu vực của họ.
* Quản lý rủi ro và bảo hiểm: Thông tin thời tiết giúp doanh nghiệp đánh giá rủi ro và đảm bảo tài sản. Đối với các ngành như nông nghiệp, việc theo dõi thời tiết giúp quyết định bảo vệ cây trồng khỏi tác động của cơn bão hoặc hạn hán.

Việc tích hợp thông tin thời tiết vào quyết định kinh doanh giúp doanh nghiệp linh hoạt hơn trong hoạt động và định hướng tối ưu hóa các chiến lược.

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

* **Đạt được:**

Sau khi kết thúc kỳ thực tập đề tài **"ÁP DỤNG MÔ HÌNH HỌC MÁY TRONG PHÂN TÍCH VÀ DỰ ĐOÁN DỮ LIỆU THỜI TIẾT"**, đã có một loạt kết quả đáng kể đạt được. Những kết quả này không chỉ mang lại giá trị trong lĩnh vực phân tích dữ liệu thời tiết, mà còn có tác động tích cực đến nhiều lĩnh vực khác trong cuộc sống và công việc.

* 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.

• Về đề tài:

- 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.

- Cài đặt thành công và sử dụng thành thạo công cụ hỗ trợ trong việc trực quan hóa dữ liệu, tìm hiểu các tính năng, cách thức xây dựng Dashboard trên phần mềm Tableau.

- Biết cách ứng dụng các khái niệm, kiến thức đã học hỏi và tìm hiểu vào bài toán dữ liệu. - Thực hiện các mô hình dự đoán với các thuật toán K-Neighbors Classifier; Decision Tree Classifier và Random Forest giúp chuyển đổi dữ liệu thành thông tin có ý nghĩa nhằm đánh giá độ tin cậy và mức độ chính xác của mô hình. Từ đó, áp dụng trong việc khai phá dữ liệu, hỗ trợ việc ra quyết định của doanh nghiệp trong hoạt động quản lý dự án.

• Về bản thân:

- Bản thân được làm việc trong môi trường doanh nghiệp chuyên nghiệp, có cơ hội tiếp cận với văn hóa doanh nghiệp và phát triển các kỹ năng của mình.

- Nâng cao được kiến thức chuyên môn, học thêm các công cụ mới.

- Tích lũy được nhiều kiến thức hữu ích và vận dụng chúng vào thực tế một cách hiệu quả.

- Bên cạnh đó, nâng cao được các kỹ năng mềm như giải quyết vấn đề phát sinh, giao tiếp với đồng nghiệp, quản lý thời gian và nhiều kỹ năng khác cần thiết cho công việc.

* **Hạn chế:**
* Bộ dữ liệu thực hiện đề tài là bộ dữ liệu mẫu, do đó các vấn đề thực tiễn khi làm việc chưa được áp dụng đầy đủ

- 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.

- Chưa triển khai được các mô hình dự đoán khác trên những công cụ khác

* **Hướng phát triển:**

• Về đề tài

- Trong thời gian tới, sẽ tiếp tục phát triển khai xây dựng mô hình khai phá dữ liệu bằng các mô hình khác để đưa ra nhiều tri thức giúp hỗ trợ tối ưu hơn trong việc ra quyết định.

• Về bản thân

- Tiếp tục tìm hiểu và học hỏi thêm nhiều về công việc Data Analyst, công việc Business Intelligence.

- 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

# TÀI LIỆU THAM KHẢO

* + - 1. GOYAL, S. (2021). Được truy lục từ <https://www.kaggle.com/datasets/vanviethieuanh/vietnam-weather-data>
      2. Việt Đà(2023): <https://www.vietda.com.vn/gioi-thieu/>
      3. Data analyst (2023): <https://chat.openai.com/c/79501523-7c0a-43ac-a956-cd6a1ac05bf7>
      4. Tableau (2023): <https://www.tableau.com/learn/get-started/dashboards>
      5. Machine Learning (2018) : <https://viblo.asia/p/machine-learning-thu-lam-nha-thien-van-du-bao-thoi-tiet-djeZ1xYmKWz>
      6. <https://insight.isb.edu.vn/data-mining-la-gi-cac-cong-cu-khai-pha-du-lieu/>
      7. <https://chat.openai.com/>