**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CƠ SỞ**

**Tên đề tài:**

**MÔ HÌNH DỰ ĐOÁN, DỰ BÁO THỜI TIẾT**

**Giảng viên hướng dẫn: ĐẶNG NHƯ PHÚ**

**Sinh viên thực hiện: PHẠM NHẬT KHÁNH**

**MSSV: 2100011313**

**Khoá: 21**

**Ngành/ chuyên ngành: TRÍ TUỆ NHÂN TẠO**

TP. HCM, tháng 9 năm 2023

**TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CƠ SỞ**

**Tên đề tài:**

**MÔ HÌNH DỰ ĐOÁN, DỰ BÁO THỜI TIẾT**

**Giảng viên hướng dẫn: ĐẶNG NHƯ PHÚ**

**Sinh viên thực hiện: PHẠM NHẬT KHÁNH**

**MSSV: 2100011313**

**Khoá: 21**

**Ngành/ chuyên ngành: TRÍ TUỆ NHÂN TẠO**

TP. HCM, tháng 9 năm 2023

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **KHOA CÔNG NGHỆ THÔNG TIN**  🙜 🙜 🙝 | CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM  **Độc lập – Tự do – Hạnh phúc**  🙜 🙜 🙝 |

# NHIỆM VỤ ĐỒ ÁN MÔN HỌC

*(Sinh viên phải đóng tờ này vào cuốn báo cáo)*

Họ và tên: **PHẠM NHẬT KHÁNH** MSSV: **2100011313**

Chuyên ngành: **Trí tuệ nhân tạo** Lớp: **21DHT1D**

Email: **NHATKHANHMEN29@GMAIL.COM** SĐT: **0385310781** Tên đề tài: MÔ HÌNH DỰ ĐOÁN, DỰ BÁO THỜI TIẾT

Giảng viên giảng dạy: **ThS. Đặng Như Phú**

Thời gian thực hiện: **26/6/2023** **đến 29/09/2023**

Nhiệm vụ/nội dung (mô tả chi tiết nội dung, yêu cầu, phương pháp… ):

## Mục tiêu

* Tìm hiểu và áp dụng các kỹ thuật cơ bản của Trí tuệ nhân tạo để dự đoán dựa trên các yếu tố của đầu vào.
* Hiểu được quy trình xây dựng mô hình, huấn luyện và đánh giá mô hình trong Trí tuệ nhân tạo.

## Phạm vi

* Thu thập dữ liệu liên quan đến Mô hình dự đoán, dự báo thời tiết.
* Tiến hành tiền xử lý dữ liệu để chuẩn bị cho việc huấn luyện mô hình.
* Xây dựng mô hình dự đoán bằng các phương pháp học máy cơ bản (ví dụ: hồi quy tuyến tính).

## Yêu cầu

1. Thu thập dữ liệu và tiền xử lý:

* Thu thập tập dữ liệu và các yếu tố liên quan đến Mô hình dự đoán, dự báo thời tiết.
* Tiến hành tiền xử lý dữ liệu để loại bỏ nhiễu và chuẩn hóa.

2. Xây dựng mô hình dự đoán:

* Xây dựng mô hình học máy cơ bản (ví dụ: hồi quy tuyến tính) để dự đoán dựa trên các yếu tố đã chọn.

3. Huấn luyện và đánh giá mô hình:

* Tách tập dữ liệu thành tập huấn luyện và tập kiểm tra.
* Huấn luyện mô hình trên tập huấn luyện và đánh giá hiệu suất trên tập kiểm tra.

4. Thực hiện ứng dụng:

* Xây dựng ứng dụng đơn giản cho phép người dùng nhập thông tin và dự đoán kết quả.

5. Báo cáo đồ án:

* Viết báo cáo về quy trình thực hiện, kết quả đạt được và đánh giá mô hình.
* Chuẩn bị thuyết trình để trình bày kết quả đồ án.

## Tiến độ dự kiến

* Tuần 1-2: Nghiên cứu lý thuyết và thu thập dữ liệu.
* Tuần 3-4: Tiền xử lý dữ liệu và xây dựng mô hình dự đoán.
* Tuần 5-6: Huấn luyện và đánh giá mô hình.
* Tuần 7-8: Xây dựng ứng dụng và tối ưu hóa.
* Tuần 9-10: Hoàn thiện báo cáo đồ án và nội dung thuyết trình.

**Nội dung và yêu cầu đã được thông qua Bộ môn.**

*Tp.HCM, ngày tháng 9 năm 2023*

|  |  |
| --- | --- |
| **TRƯỞNG BỘ MÔN**  *(Ký và ghi rõ họ tên)*  **ThS. Đặng Như Phú** | **GIẢNG VIÊN GIẢNG DẠY**  *(Ký và ghi rõ họ tên)*  **ThS. Đặng Như Phú** |

# LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Đại học Nguyễn Tất Thành, đặc biệt là thầy Đặng Như Phú đã tận tình hướng dẫn cho em những thông tin, kiến thức, kĩ năng cần thiết để có thể hoàn thành bài báo cáo trên.

Trong quá trình tìm hiểu, thực hiện khó tránh khỏi những sai sót, kính mong các thầy cô bỏ qua. Đồng thời do trình độ lý luận cũng như kĩ năng và kiến thức còn hạn chế nên bài báo cáo không thể tránh khỏi những thiếu sót. Em rất mong nhận được những ý kiến của các thầy cô để em học thêm được nhiều kinh nghiệm và sẽ hoàn thành tốt hơn trong những bài báo cáo sau.

Sau cùng em xin kính chúc quý thầy cô dồi dào sức khỏe, niềm tin để tiếp tục thực hiện sứ mệnh cao đẹp truyền đạt kiến thức cho thế hệ mai sau và luôn thành công tốt đẹp trong công việc.

# LỜI MỞ ĐẦU

Thời tiết luôn là yếu tố quan trọng trong cuộc sống hằng ngày của chúng ta. Dự báo thời tiết chính xác sẽ là bước đẩy quan trọng giúp cho chúng ta tự chủ trong mọi hành động, tránh những rủi ro xấu có thể xảy ra bất cứ lúc nào. Thay vì dự báo thời tiết qua những kinh nghiệm dân gian xưa như: quan sát hình dạng đám mây và hướng di chuyển của chúng, quan sát trăng và sao hay quan sát cách di chuyển đường bay của chuồn chuồn… Mô hình này sẽ thông qua các thông số được cập nhật sẽ chuẩn đoán ra thời tiết hôm nay sẽ như thế nào, từ đó giúp mọi người chủ động hơn trong công việc, hoạt động sinh hoạt đời thường, đặc biệt là trong thị trường hàng không, sẽ hạn chế các rủi ro cho những chuyến bay tránh thời tiết xấu, đem đến cho khách hàng sự an tâm, đồng thời phát triển nền kinh tế hàng không nói riêng và nền kinh tế đất nước nói chung.

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ …..… NĂM HỌC …….. - ….…** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

BM-ChT-11

Môn thi: …………………………………………… Lớp học phần:……………………………

Nhóm sinh viên thực hiện:…………………………

1……………………………………. Tham gia đóng góp: …………………….

2……………………………………. Tham gia đóng góp: …………………….

3……………………………………. Tham gia đóng góp: …………………….

4……………………………………. Tham gia đóng góp: …………………….

5……………………………………. Tham gia đóng góp: …………………….

6……………………………………. Tham gia đóng góp: …………………….

7……………………………………. Tham gia đóng góp: …………………….

8……………………………………. Tham gia đóng góp: …………………….

Ngày thi:…………………………………………… Phòng thi: ……………………………….

Đề tài tiểu luận/Báo cáo của sinh viên:………………………………………………………….

…………………………………………………………………………………………………...

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo | ………………………………………………………  ……………………………………………………… | 1 |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần | ………………………………………………………  ……………………………………………………… | 6 |  |
| * Lập luận | ………………………………………………………  ……………………………………………………… | 2 |  |
| * Kết luận | ……………………………………………………… | 0.5 |  |
| Trình bày | ……………………………………………………… | 0.5 |  |
| **TỔNG ĐIỂM** |  | **10** |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)*  **Đặng Như Phú** |

# MỤC LỤC

[LỜI CẢM ƠN i](#_Toc146841591)

[LỜI MỞ ĐẦU ii](#_Toc146841592)

[MỤC LỤC iv](#_Toc146841593)

[DANH MỤC HÌNH v](#_Toc146841594)

[KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT vi](#_Toc146841595)

[CHƯƠNG 1 1](#_Toc146841596)

[CƠ SỞ LÝ THUYẾT 1](#_Toc146841597)

[1.1. Học máy (Machine learning) 1](#_Toc146841598)

[1.2 Các loại thuật toán học máy 2](#_Toc146841599)

[1.3 Ứng dụng 3](#_Toc146841600)

[1.4 Một số mô hình trong Machine learning 3](#_Toc146841601)

[CHƯƠNG 2 6](#_Toc146841602)

[PHÂN TÍCH YÊU CẦU 6](#_Toc146841603)

[2.1. Thu thập dữ liệu và tiền xử lý 6](#_Toc146841604)

[2.1.1 Thu thập dữ liệu liên quan đến mô hình 6](#_Toc146841605)

[2.1.2 Tiền xử lý dữ liệu 6](#_Toc146841606)

[2.2 Mô tả dữ liệu thu thập được 7](#_Toc146841607)

[CHƯƠNG 3 8](#_Toc146841608)

[XÂY DỰNG MÔ HÌNH 8](#_Toc146841609)

[3.1. Xây dựng mô hình dự đoán 8](#_Toc146841610)

[3.1.1 Các giải thuật sử dụng để xây dựng mô hình dự báo thời tiết 8](#_Toc146841611)

[3.1.2 Quá trình xây dựng mô hình dự báo thời tiết 9](#_Toc146841612)

[3.1.2.1 Khai báo các thư viện 9](#_Toc146841613)

[3.1.2.2 Tải lên dữ liệu đầu vào và đọc dữ 11](#_Toc146841614)

[3.1.2.3 Kiểm tra và hiển thị dữ liệu dưới dạng biểu đồ 12](#_Toc146841615)

[3.1.2.4 Lấy ra dữ liệu cần thiết 13](#_Toc146841616)

[3.1.2.5 Huấn luyện mô hình 13](#_Toc146841617)

[CHƯƠNG 4 16](#_Toc146841618)

[THỰC NGHIỆM MÔ HÌNH 16](#_Toc146841619)

[4.1. Dự đoán 16](#_Toc146841620)

[4.2 Đánh giá mô hình 16](#_Toc146841621)

[CHƯƠNG 5 18](#_Toc146841622)

[KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 18](#_Toc146841623)

[5.1. Kết quả đạt được 18](#_Toc146841624)

[5.2. Hạn chế của đề tài 18](#_Toc146841625)

[5.3. Hướng phát triển 18](#_Toc146841626)

[TÀI LIỆU THAM KHẢO 20](#_Toc146841627)

# DANH MỤC HÌNH

[*Hình 1. 1- Học máy có giám sát* 2](#_Toc146754384)

[*Hình 1. 3- Học máy không giám sát* 3](#_Toc146754385)

[*Hình 2. 1- Dữ liệu mẫu từ Kaggle* 6](#_Toc146754393)

[*Hình 2. 2- Mô tả dữ liệu thu thập được* 7](#_Toc146754394)

[*Hình 3. 1 Khai báo các thư viện cần sử dụng*  9](#_Toc146754398)

[*Hình 3. 2- Tải file* 11](#_Toc146754399)

[*Hình 3. 3- Đọc file* 11](#_Toc146754400)

[*Hình 3. 4- Hiển thị dữ liệu* 12](#_Toc146754401)

[*Hình 3. 5- Dữ liệu thể hiện dưới dạng biểu đồ* 12](#_Toc146754402)

[*Hình 3. 6- Lấy dữ liệu cần thiết* 13](#_Toc146754403)

[*Hình 3. 7- Tách dữ liệu train và test* 13](#_Toc146754404)

[*Hình 3. 8- Chuẩn hoá dữ liệu* 13](#_Toc146754405)

[*Hình 3. 9- Train cho AI* 14](#_Toc146754406)

[*Hình 3. 10- Lấy kết quả dự đoán* 14](#_Toc146754407)

[*Hình 3. 11- So sánh kết quả* 14](#_Toc146754408)

[*Hình 3. 12- In điểm của 3 thuật toán* 15](#_Toc146754409)

[*Hình 4. 1- Thực nghiệm mô hình* 16](#_Toc146754410)

# KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT

|  |  |
| --- | --- |
| **Chữ viết tắt** | **Ý nghĩa** |
| CSDL | Cơ sở dữ liệu |
| ML | Machine Learning |

# CHƯƠNG 1

# CƠ SỞ LÝ THUYẾT

## 1.1. Học máy (Machine learning)

Học máy hay máy học trong tiếng Anh là Machine learning, viết tắt: ML.

Học máy (ML) là một công nghệ phát triển từ lĩnh vực trí tuệ nhân tạo. Các thuật toán ML là các chương trình máy tính có khả năng học hỏi về cách hoàn thành các nhiệm vụ và cách cải thiện hiệu suất theo thời gian.

ML vẫn đòi hỏi sự đánh giá của con người trong việc tìm hiểu dữ liệu cơ sở và lựa chọn các kĩ thuật phù hợp để phân tích dữ liệu. Đồng thời, trước khi sử dụng, dữ liệu phải sạch, không có sai lệch và không có dữ liệu giả.

Các mô hình ML yêu cầu lượng dữ liệu đủ lớn để "huấn luyện" và đánh giá mô hình. Trước đây, các thuật toán ML thiếu quyền truy cập vào một lượng lớn dữ liệu cần thiết để mô hình hóa các mối quan hệ giữa các dữ liệu. Sự tăng trưởng trong dữ liệu lớn (big data) đã cung cấp các thuật toán ML với đủ dữ liệu để cải thiện độ chính xác của mô hình và dự đoán.



*Hình 1. 1- Học máy*

## 1.2 Các loại thuật toán học máy

Có 2 loại ML chính bao gồm học có giám sát (supervised learning) và học không giám sát (unsupervised learning).

**🡪 Học máy có giám sát**

Trong học có giám sát, máy tính học cách mô hình hóa các mối quan hệ dựa trên dữ liệu được gán nhãn (labeled data). Sau khi tìm hiểu cách tốt nhất để mô hình hóa các mối quan hệ cho dữ liệu được gắn nhãn, các thuật toán được huấn luyện được sử dụng cho các bộ dữ liệu mới.

Ứng dụng của kĩ thuật học có giám sát: Xác định tín hiệu hay biến số tốt nhất để dự báo lợi nhuận trong tương lai của cổ phiếu hoặc dự đoán xu hướng thị trường chứng khoán.

A diagram of a diagram

Description automatically generated

*Hình 1. 1- Học máy có giám sát*

**🡪Học máy không giám sát**

Trong học không giám sát, máy tính không được cung cấp dữ liệu được dán nhãn mà thay vào đó chỉ được cung cấp dữ liệu mà thuật toán tìm cách mô tả dữ liệu và cấu trúc của chúng.

Ứng dụng của học không giám sát: Phân loại các công ty thành các nhóm công ty tương đồng dựa trên đặc điểm của chúng thay vì sử dụng tiêu chuẩn của các nhóm ngành hoặc các quốc gia.

**A diagram of a machine learning

Description automatically generated**

*Hình 1. 2- Học máy không giám sát*

## 1.3 Ứng dụng

Các thuật toán ML đang được sử dụng để phân tích dữ liệu lớn (big data) để giúp dự đoán xu hướng hoặc sự kiện thị trường, ví dụ như dự đoán kết quả cuộc bầu cử chính trị.

Các thuật toán nhận dạng hình ảnh hiện có thể phân tích dữ liệu từ các hệ thống chụp ảnh vệ tinh để cung cấp thông tin về số lượng khách hàng tại các bãi đậu xe của cửa hàng bán lẻ, hoạt động vận chuyển và cơ sở sản xuất, và sản lượng nông nghiệp...Những thông tin này sẽ cung cấp dữ liệu đầu vào cho các mô hình định giá hoặc các mô hình kinh tế.

## 1.4 Một số mô hình trong Machine learning

**🡪Mô hình** **Naive Bayes**

Mô hình Naive Bayes trong Machine Learning là một phương pháp phân loại dựa trên nguyên tắc của định lý Bayes. Nó được gọi là "naive" (ngây thơ) vì giả định rằng các đặc trưng đầu vào là độc lập tuyến tính, tức là không có sự tương quan hoặc ảnh hưởng lẫn nhau.

Mô hình Naive Bayes tính toán xác suất của một mẫu thuộc về mỗi lớp dựa trên xác suất tiền đề và xác suất hậu nghiệm. Nó sử dụng công thức Bayes để tính toán xác suất hậu nghiệm dựa trên xác suất tiền đề và xác suất điều kiện. Mô hình này rất phù hợp cho các bài toán phân loại văn bản, phân loại email spam, phân loại tin tức, và nhiều bài toán khác.

Mô hình Naive Bayes có hiệu suất tốt và tốc độ huấn luyện nhanh, đặc biệt là khi số lượng đặc trưng lớn. Tuy nhiên, giả định về tính độc lập tuyến tính của các đặc trưng có thể không phù hợp với mọi bài toán thực tế.

**🡪Mô hình Logistic regression**

Mô hình Logistic Regression (Hồi quy Logistic) trong Machine Learning là một phương pháp được sử dụng để dự đoán và phân loại dữ liệu dựa trên xác suất. Mặc dù có từ "regression" trong tên gọi, nhưng Logistic Regression thực chất là một phương pháp phân loại, không phải hồi quy dự đoán giá trị liên tục.

Mô hình Logistic Regression sử dụng hàm logistic (hay còn gọi là hàm sigmoid) để biểu diễn xác suất thuộc về mỗi lớp. Hàm sigmoid có dạng S-shaped và chuyển đổi giá trị đầu vào thành một giá trị xác suất nằm trong khoảng từ 0 đến 1. Mô hình sử dụng các tham số để tìm một đường cong tối ưu để phân loại dữ liệu.

Quá trình huấn luyện mô hình Logistic Regression thường được thực hiện bằng cách tối ưu hóa hàm mất mát (loss function) thông qua các phương pháp như Gradient Descent. Mục tiêu là tìm ra các tham số tối ưu để mô hình có thể dự đoán xác suất chính xác nhất cho từng lớp dữ liệu.

**🡪Mô hình Decision Tree**

Mô hình Decision Tree (Cây quyết định) trong Machine Learning là một phương pháp học có giám sát được sử dụng cho bài toán phân loại và dự đoán. Nó dựa trên cấu trúc cây để tạo ra các quyết định dựa trên các đặc trưng của dữ liệu đầu vào.

Cây quyết định bắt đầu từ một nút gốc và tiếp tục phân chia dữ liệu thành các nhánh con dựa trên các đặc trưng. Quá trình phân chia được thực hiện bằng cách đặt các câu hỏi về các đặc trưng và dựa trên câu trả lời để di chuyển từ nút gốc đến các nút lá. Mỗi nút lá đại diện cho một lớp hoặc một giá trị dự đoán.

Quá trình xây dựng cây quyết định thường dựa trên các thuật toán như ID3, C4.5 hoặc CART. Các thuật toán này sẽ đánh giá các đặc trưng và lựa chọn đặc trưng tốt nhất để phân chia dữ liệu. Mục tiêu là tạo ra một cây quyết định có khả năng phân loại tốt và dễ hiểu.

Mô hình Decision Tree có nhiều ưu điểm, bao gồm khả năng xử lý dữ liệu số, dữ liệu rời rạc và dữ liệu có giá trị bị thiếu. Nó cũng có khả năng giải thích kết quả dễ dàng và không yêu cầu nhiều tiền xử lý dữ liệu. Tuy nhiên, cây quyết định có thể dễ dàng bị overfitting nếu không được kiểm soát, và có thể không tạo ra kết quả tốt khi dữ liệu phức tạp.

# CHƯƠNG 2

# PHÂN TÍCH YÊU CẦU

## 2.1. Thu thập dữ liệu và tiền xử lý

### 2.1.1 Thu thập dữ liệu liên quan đến mô hình

Dữ liệu là một yếu tố bắt buộc trong tất cả các bài toán Machine Learning. Với bài toán dự báo thời tiết dựa trên thông số cho trước, có thể tìm kiếm và download dữ liệu tại trang Kaggle để phục vụ cho việc tiến hành thu thập dữ liệu và khởi chạy chương trình.

Dữ liệu cần thu thập là một số thông tin về thời tiết, thông số, ngày tháng ghi chép dữ liệu, cụ thể:

- Thời điểm ghi chép dữ liệu: ngày tháng cụ thể của từng mẫu thời tiết được ghi lại rõ ràng, có thứ tự.

- Độ ngưng tụ hơi nước được đánh giá dựa trên thang đo từ mức 0.

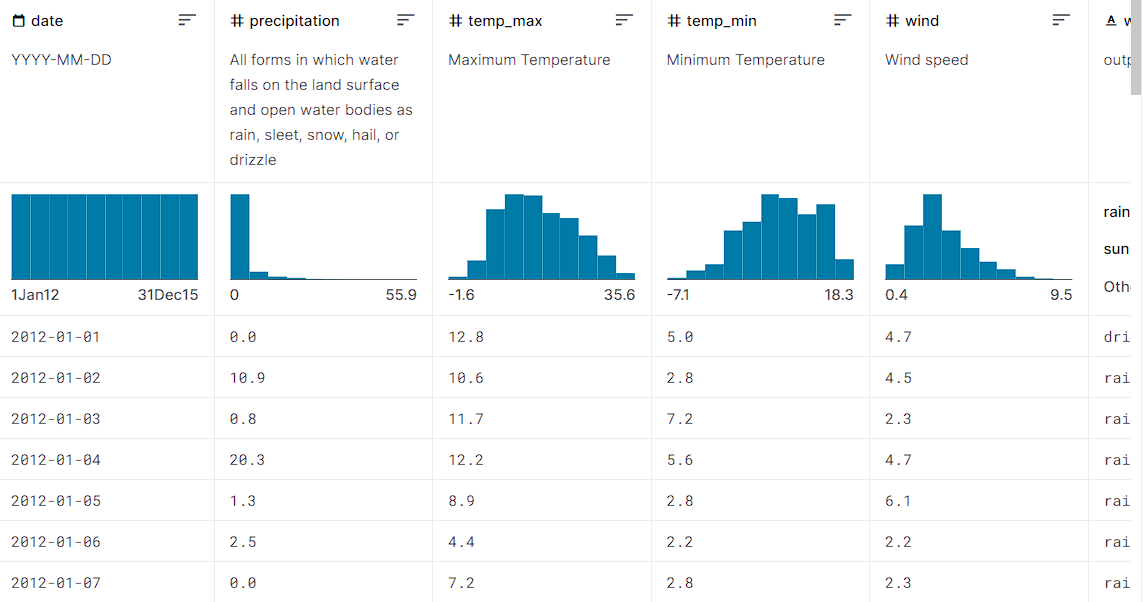
- Nhiệt độ (chia làm nhiệt độ cao nhất và nhỏ nhất): nhiệt độ được đo tại các trạm khí tượng cung cấp số liệu báo cáo, đơn vị đo ºC.

- Sức gió.

- Trạng thái thời tiết.

### 2.1.2 Tiền xử lý dữ liệu

Sau khi tiến hành thu thập dữ liệu, em đã kiểm tra tổng quan dữ liệu và thấy số liệu ổn định, không có sự rối loạn giữa các số liệu. Do dữ liệu em thu thập là ở dạng số, nên không cần chuẩn hoá dữ liệu về dạng số.



*Hình 2. 1- Dữ liệu mẫu từ Kaggle*

## 2.2 Mô tả dữ liệu thu thập được

- Thuộc tính mô tả: bao gồm loại dữ liệu và kiểu dữ liệu

+ Thời gian ghi dữ liệu (date): object

+ Độ ngưng tụ hơi nước (precipitation): float64

+ Nhiệt độ cao nhất (temp\_max): float64

+ Nhiệt độ thấp nhất (temp\_min): float 64

+ Sức gió (wind): float 64

- Thuộc tính nhãn (weather): bao gồm loại thời tiết và kiểu dữ liệu

+ Mưa (rain): object

+ Nắng (sun): object

+ Tuyết (snow): object

+ Mưa phùn (drizzle): object

+ Sương mù (fog): object

A screenshot of a computer code

Description automatically generated

*Hình 2. 2- Mô tả dữ liệu thu thập được*

# CHƯƠNG 3

# XÂY DỰNG MÔ HÌNH

## 3.1. Xây dựng mô hình dự đoán

### 3.1.1 Các giải thuật sử dụng để xây dựng mô hình dự báo thời tiết

**🡪 Giải thuật Naïve Bayes**

Giải thuật Naive Bayes là một phương pháp học máy dựa trên lý thuyết xác suất. Nó được sử dụng rộng rãi trong phân loại và dự đoán. Giải thuật này dựa trên định lý Bayes để tính toán xác suất của một sự kiện dựa trên các thông tin có sẵn.

Cách hoạt động của Naive Bayes là xác định xác suất của một mục tiêu dựa trên các đặc trưng đầu vào. Nó giả định rằng các đặc trưng đầu vào là độc lập tuyến tính, điều này thường không đúng trong thực tế. Mặc dù giả định này khá đơn giản, Naive Bayes vẫn cho kết quả tốt trong nhiều trường hợp thực tế.

Để tính toán xác suất, Naive Bayes sử dụng công thức Bayes:

P(A|B) = (P(B|A) \* P(A)) / P(B)

Trong đó:

- P(A|B) là xác suất của sự kiện A xảy ra khi biết sự kiện B đã xảy ra.

- P(B|A) là xác suất của sự kiện B xảy ra khi biết sự kiện A đã xảy ra.

- P(A) và P(B) là xác suất của sự kiện A và B xảy ra độc lập.

Naive Bayes được áp dụng trong nhiều lĩnh vực, bao gồm phân loại văn bản, phân loại email spam, phân loại ảnh, và nhiều ứng dụng khác.

**🡪 Giải thuật Logistic Regression**

Giải thuật Logistic Regression là một phương pháp học máy được sử dụng để dự đoán và phân loại dữ liệu. Nó dựa trên mô hình hồi quy tuyến tính và sử dụng hàm sigmoid để ánh xạ đầu ra thành một xác suất.

Cách hoạt động của Logistic Regression là tìm một đường cong tuyến tính để phân tách các điểm dữ liệu thành các nhóm khác nhau. Đầu tiên, dữ liệu được biểu diễn bằng các đặc trưng đầu vào và nhãn đầu ra. Sau đó, một hàm sigmoid được áp dụng để tính toán xác suất dự đoán cho mỗi điểm dữ liệu. Hàm sigmoid chuyển đổi giá trị đầu ra thành một giá trị nằm trong khoảng từ 0 đến 1, đại diện cho xác suất thuộc vào một nhóm cụ thể.

Tiếp theo, giải thuật sử dụng phương pháp tối ưu hóa để điều chỉnh các tham số của đường cong tuyến tính, nhằm tối thiểu hóa sai số giữa dự đoán và nhãn thực tế. Phương pháp tối ưu hóa thường là Gradient Descent, trong đó giải thuật cập nhật các tham số dựa trên độ dốc của hàm mất mát.

Cuối cùng, sau khi huấn luyện mô hình, nó có thể được sử dụng để dự đoán nhãn cho các điểm dữ liệu mới dựa trên đặc trưng đầu vào.

**🡪 Giải thuật Decision Tree (cây quyết định)**

Giải thuật Decision Tree (cây quyết định) là một phương pháp học máy được sử dụng để phân loại và dự đoán dữ liệu. Nó dựa trên việc xây dựng một cây quyết định từ dữ liệu huấn luyện, trong đó mỗi nút trong cây đại diện cho một quyết định hoặc một thuộc tính để phân loại dữ liệu.

Cách hoạt động của Decision Tree là chia dữ liệu thành các nhánh dựa trên các thuộc tính và giá trị của chúng. Quá trình chia dữ liệu được thực hiện dựa trên các tiêu chí như độ tinh khiết của nhánh hoặc độ tương đồng giữa các điểm dữ liệu trong cùng một nhánh. Quá trình này tiếp tục cho đến khi không còn cách chia nào tốt hơn hoặc đạt được một điều kiện dừng.

Khi cây quyết định đã được xây dựng, nó có thể được sử dụng để phân loại dữ liệu mới bằng cách đi qua các quyết định từ gốc cây đến lá tương ứng với kết quả phân loại. Decision Tree có thể xử lý cả dữ liệu rời rạc và dữ liệu liên tục và thường dễ hiểu và giải thích.

### 3.1.2 Quá trình xây dựng mô hình dự báo thời tiết

#### 3.1.2.1 Khai báo các thư viện

**A screenshot of a computer

Description automatically generated**

*Hình 3. 1 Khai báo các thư viện cần sử dụng*

- Các thư viện và hàm được sử dụng:

**+ NumPy**: Thư viện NumPy là một thư viện Python mạnh mẽ và phổ biến được sử dụng cho tính toán khoa học và toán học. NumPy cung cấp một cấu trúc dữ liệu mảng nhiều chiều (ndarray) linh hoạt và hiệu suất cao, cùng với một loạt các chức năng và phương pháp để làm việc với mảng này.

**+ Pandas**: Khai báo thư viện hỗ trợ đọc và phân tích dữ liệu ở dạng bảng,đa chiều, có tiềm năng không đồng nhất) và dữ liệu chuỗi thờigian.

**+ Matplotlib**: Là một thư viện Python sử dụng Python Script để giúp tạo racác đồ thị 2D. Thư viện này có hỗ trợ tạo nhiều giao điểm giữahai trục số trong cùng một lúc. Ngoài ra cũng có thể dung Matplotlib để thao tác trực tiếp đến các đặc điểm khác nhau của đồ thị.

**+ Seaborn**: Là một trong những thư viện Python được đánh giá cao nhấtthế giới được xây dựng nhằm mục đích tạo ra các hình ảnh trựcquan đẹp mắt. Nó có thể được coi là một phần mở rộng củamột thư viện khác có tên là Matplotlib vì nó được xây dựngtrên đó.

**+ Sklearn:** Là thư viện mạnh mẽ nhất dành cho các thuật toán học máyđược viết trên ngôn ngữ Python. Thư viện cung cấp một tập cáccông cụ xử lý các bài toán machine learning và statisticalmodeling gồm: classification, regression, clustering, và 12 dimensionality reduction.

**+ train\_test\_split**: Là một hàm trong thư viện scikit-learn của Python, được sử dụng để chia dữ liệu thành hai tập con: tập huấn luyện (training set) và tập kiểm tra (test set).

**+** **GaussianNB**: Là một mô hình học máy trong thư viện scikit-learn của Python, được sử dụng cho bài toán phân loại. Nó dựa trên giả định rằng các đặc trưng đầu vào tuân theo phân phối Gaussian (phân phối chuẩn).

**+** **DecisionTreeClassifier**: Là một mô hình học máy trong thư viện scikit-learn của Python, được sử dụng cho bài toán phân loại. Nó dựa trên cấu trúc cây quyết định để tạo ra các quy tắc phân loại dựa trên các đặc trưng của dữ liệu.

**+ Logistic Regression**: Được sử dụng để dự đoán xác suất của một mẫu dữ liệu thuộc vào một lớp cụ thể. Nó sử dụng hàm sigmoid (hàm logistic) để biểu diễn xác suất dự đoán và áp dụng một ngưỡng để quyết định lớp cuối cùng của mẫu dữ liệu. loại.

**+** **StandardScaler**: Được sử dụng để chuẩn hóa các đặc trưng của dữ liệu sao cho chúng có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Quá trình chuẩn hóa này giúp đảm bảo rằng các đặc trưng có cùng phạm vi giá trị và không bị ảnh hưởng bởi đơn vị đo lường ban đầu.

**+** **accuracy\_score**: Là một phương pháp để tính toán tỷ lệ phân loại chính xác của mô hình. Nó được tính bằng cách so sánh các dự đoán của mô hình với nhãn thực tế của dữ liệu và tính toán tỷ lệ phân loại chính xác. Kết quả của accuracy\_score là một giá trị nằm trong khoảng từ 0 đến 1, với 1 đại diện cho 100% phân loại chính xác.

**+** **confusion\_matrix**: Là một ma trận được sử dụng để đánh giá hiệu suất của mô hình phân loại. Nó hiển thị số lượng dự đoán đúng và sai lệch cho từng lớp trong bài toán phân loại. Ma trận này có kích thước bằng số lượng lớp trong bài toán phân loại. Bằng cách xem xét các giá trị trong ma trận, ta có thể đánh giá được mức độ phân loại chính xác và các loại sai lệch (như false positive và false negative) của mô hình.

#### 3.1.2.2 Tải lên dữ liệu đầu vào và đọc dữ

Sau khi tải file data từ trang Kaggle về máy ta cần thực hiện bước sau đây để tải dữ liệu lên Google Colab

A screenshot of a computer

Description automatically generated

*Hình 3. 2- Tải file*

Để đọc file dữ liệu trên ta thực hiện:



*Hình 3. 3- Đọc file*

#### 3.1.2.3 Kiểm tra và hiển thị dữ liệu dưới dạng biểu đồ

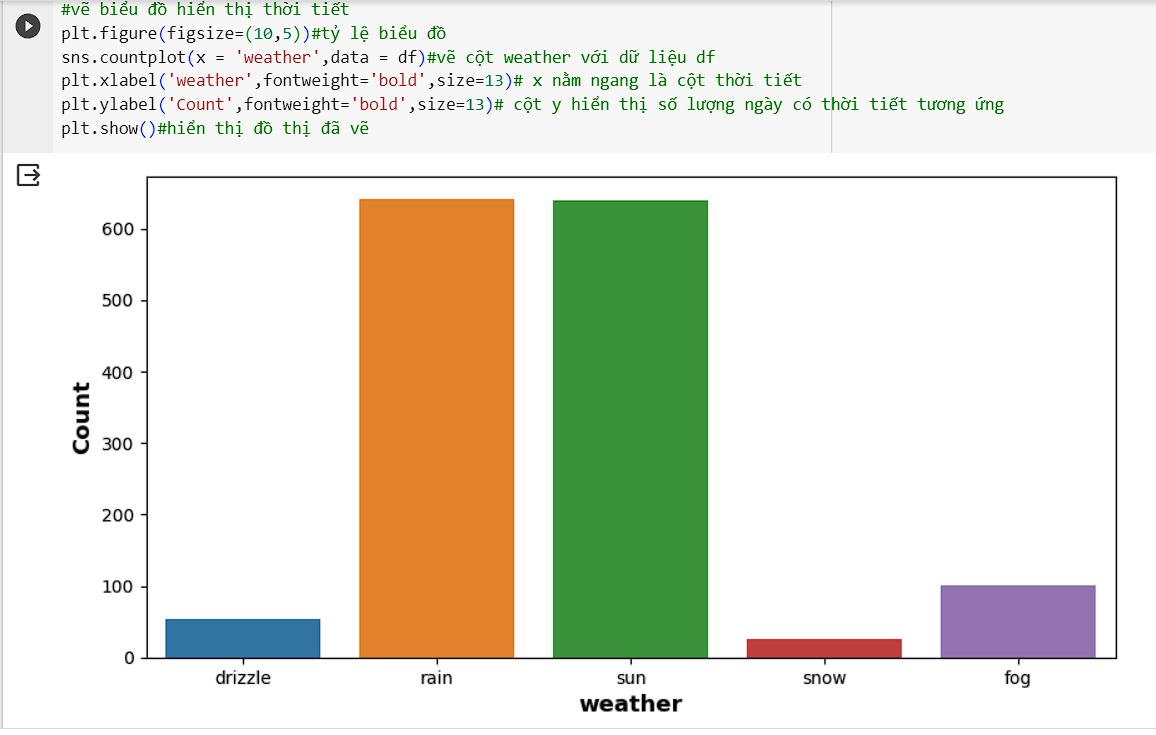
Kiểm tra dữ liệu đầu vào đã đúng như ta mong đợi hay chưa ta thực hiện:

A screenshot of a computer

Description automatically generated

*Hình 3. 4- Hiển thị dữ liệu*

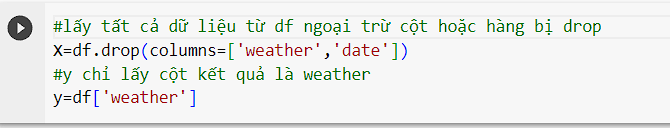
Hiển thị dữ liệu dưới dạng biểu đồ sau đây



*Hình 3. 5- Dữ liệu thể hiện dưới dạng biểu đồ*

Biểu đồ cho ta thấy được số lượng ngày có mưa và có nắng trên 600 ngày, số lượng ngày có sương khoảng 100 ngày và số lượng ngày có tuyết và mưa phùn dưới 100 ngày.

#### 3.1.2.4 Lấy ra dữ liệu cần thiết



*Hình 3. 6- Lấy dữ liệu cần thiết*

Trong bảng dữ liệu gốc bao gồm cả dữ liệu điều kiện (Precipitation,temp\_max, temp\_min, wind) và kết quả (weather).

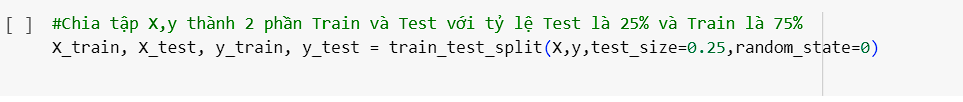
Để dự đoán ra được kết quả ta sẽ sử dụng phần dữ liệu điều kiện để dự đoán nên ta sử dụng lệnh.

x = df.drop(colums = [‘weather’,’date’] tức là lấy tất cả df ngoại trừ cộ thoặc hàng weather và date.

Sau đó chỉ lấy cột kết quả ‘Weather’ là vì cột date không ảnh hưởng đến kết quả mưa hay nắng...) để loại trừ phần kết quả ra.

#### 3.1.2.5 Huấn luyện mô hình

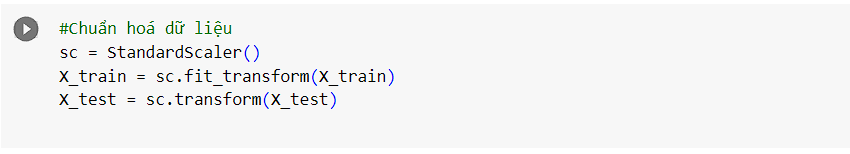
Tách dữ liệu thành tập huấn luyện và tập kiểm tra



*Hình 3. 7- Tách dữ liệu train và test*

Dữ liệu bao gồm các giá trị của cột weather.

Dữ liệu chia làm 2 phần: huấn luận 75% và kiểm tra 25%

🡪 **Chuẩn hoá dữ liệu**: bước này giúp cho giảm được lượng dữ liệu dư thừa, đảm bảo độc lập dữ liệu, giảm thiểu không gian lưu trữ dữ liệu,… giúp xử lý dữ liệu nhanh chóng hơn.

*Hình 3. 8- Chuẩn hoá dữ liệu*

StandarScaler: là một phương pháp chia tỷ lệ dựa trên trung bình. Côngthức của StandardScaler là (Xi-Xmean) / Xstd, vì vậy nó điều chỉnh giá trị trung bình là 0. StandardScaler dễ bị tác động ảnh hưởng bởi những ngoại lệ vì những ngoại lệ ảnh hưởng tác động đến giá trị trung bình.

🡪**Huấn luyện cho mô hình**: để mô hình học các dữ liệu train, ta sử dụng 3 mô hình khác nhau là Bayes, Logic và Tree. Đưa ra bộ dữ liệu điều kiện và có sẵn kết quả tương ứng để mô hình học và nhận biết.

A screenshot of a computer program

Description automatically generated

*Hình 3. 9- Train cho AI*

Sau khi train xong, đưa ra bộ dữ liệu điều kiện khác để tiến hành kiểm tra kết quả dự đoán của 3 mô hình.

ypred là dự đoán của ytest của từng mô hình.

A close-up of a test

Description automatically generated

*Hình 3. 10- Lấy kết quả dự đoán*

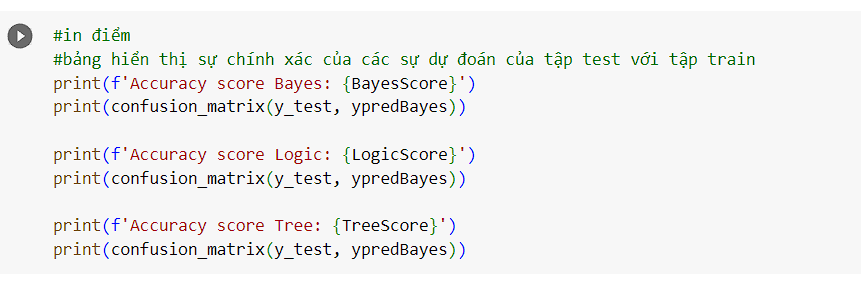
🡪**Đánh giá hiệu xuất trên tập kiểm tra:** Sau khi tiến hành kiểm tra, mô hình sẽ đưa ra kết quả dự đoán ypred. So sánh sự trùng hợp giữa ypred và ytest để đánh giá được độ chính xác của từng mô hình. Từ đó chấm điểm chính xác đối với từng mô hình.

A close-up of a code

Description automatically generated

*Hình 3. 11- So sánh kết quả*

In ra điểm độ chính xác của từng mô hình



*Hình 3. 12- In điểm của 3 thuật toán*

Ta được kết quả như sau:

A screenshot of a computer

Description automatically generated

Kết quả trên thể hiện: Hàng là giá trị ypred (giá trị bot dự đoán), cột là giá trị ytest (giá trị thực tế). Theo thứ tự từ trái sang phải, các loại thời tiết lần lượt ứng với các cột là là: drizzle, fog, sun,snow, rain. Tương tự, theo thứ tự từ trên xuống dưới, các loại thời tiết lần lượt ứng với các hàng là: drizzle, fog, sun, snow, rain. Như vậy có thể thấy, vị trí các ô bot đoán đúng (ytest và ypred trùng nhau) chính là hàng chéo của bảng trên, giá trị của các ô trong hàng chéo là số lần bot đoán đúng. Vậy các giá trị còn lại trong bảng chính là số lần bot đoán sai.

**Kết luận:**

Giá trị các ô trong hàng chéo khác 0 càng nhiều, thì mô hình đó có tỉ lệ dự đoán đúng càng cao.

# CHƯƠNG 4

# THỰC NGHIỆM MÔ HÌNH

## 4.1. Dự đoán

Cho phép người dung nhập vào các giá trị lần lượt là: độ ngưng tụ hơi nước, nhiệt độ cao nhất, nhiệt độ thấp nhất và sức gió. Sau đó kiểm tra kết quả dự đoán.

A screenshot of a computer code

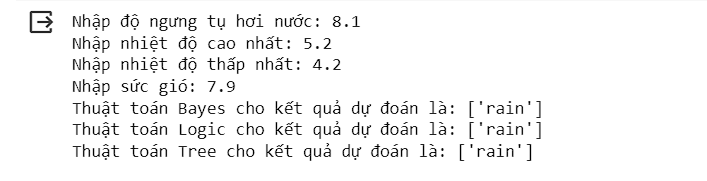
Description automatically generated

*Hình 4. 1- Thực nghiệm mô hình*

Gán testPred là một mảng chứa các giá trị cần để dự đoán là var1, var2, var3 và var4 cho bộ 3 mô hình dự đoán.

Kết quả ta sẽ được phần dự đoán thời tiết ngày hôm đó nếu các điều kiện là

các dữ liệu đưa vào ngẫu nhiên đó.



*Hình 4. 2 Kết quả dự đoán*

## 4.2 Đánh giá mô hình

**🡪Ưu điểm chung:** Cả 3 mô hình xử khá nhanh do dữ liệu đã được chuẩn hóa, đưa ra kết quả khá là sát so với thực tế với tỉ lệ đúng của cả 3 mô hình lên đến hơn 70%.

- Mô hình Naïve Bayes: Có tỉ lệ đúng cao nhất trong 3 mô hình.

- Mô hình Decision Tree: Các thuộc tính nhãn ít dữ liệu vẫn có khả năng dự đoán đúng.

**🡪Nhược điểm chung:** Dữ liệu đầu vào thực tế do có sự thiết hụt và chênh lệnh quá lớn giữa các thuộc tính nhãn nên việc sai sót là không thể tránh khỏi. Tỉ lệ dự đoán sai của cả 3 mô hình lên các thuộc tính nhãn này là khá cao.

- Mô hình Naïve Bayes: Những thuộc tính nhãn có ít dữ liệu thì dự đoán gần như là sai hết.

- Mô hình Decision Tree: Tỉ lệ đúng thấp nhất trong cả 3 mô hình.

- Mô hình Logistic regression: Những thuộc tính nhãn có ít dữ liệu thì không dự đoán đúng được cái nào cả.

# CHƯƠNG 5

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## 5.1. Kết quả đạt được

Xây dựng được một mô hình dự đoán, dự báo thời tiết có thể đoán được phần nào thời tiết sắp tới nhờ đó người dùng có thể lên kế hoạch cho các hoạt động ngoài trời, chuẩn bị cho các biện pháp bảo vệ môi trường và đảm bảo an toàn trong cuộc sống hàng ngày. Bên cạnh đó, nó còn mang lại giá trị lớn trong nhiều lĩnh vực, bao gồm nông nghiệp, hàng hải, quản lý tài nguyên tự nhiên và cả cuộc sống hàng ngày của chúng ta.

## 5.2. Hạn chế của đề tài

Là một sinh viên, nên em chưa đủ kiến thức chuyên môn để có thể xây dựng mô hình dự đoán thời tiết chính xác. Một số vấn đề mà em gặp phải:

- Hiểu biết hạn chế về lĩnh vực thời tiết.

- Thiếu dữ liệu và tài nguyên.

- Độ phức tạp cả mô hình lớn hơn so với độ hiểu biết của bản than em.

## 5.3. Hướng phát triển

Trong tương lai khi em đã đủ kiến thức chuyên môn và hiểu biết thì có thể phát triển mô hình theo các hướng sau:

Mở rộng độ phủ địa lý: Mô hình hiện tại có thể dự đoán và dự báo thời tiết cho một số vùng địa lý cụ thể. Tuy nhiên, để mở rộng ứng dụng và độ chính xác của mô hình, có thể nghiên cứu mở rộng độ phủ địa lý bằng cách bổ sung dữ liệu đa kênh từ các trạm thời tiết và cảm biến khác nhau, đồng thời cải thiện quá trình chuyển đổi dữ liệu địa lý hiện tại.

Kết hợp dữ liệu phụ trợ: Mô hình có thể được cải thiện bằng cách kết hợp dữ liệu thời tiết với các dữ liệu phụ trợ, chẳng hạn như dữ liệu khí tượng không gian, đặc điểm địa hình, thông tin về điều kiện đô thị và dữ liệu lưu lượng giao thông. Bằng cách sử dụng các dữ liệu phụ trợ này, mô hình có thể tăng cường sự hiểu biết về tương quan, tương tác và ảnh hưởng của các yếu tố bên ngoài lên thời tiết.

Sử dụng kỹ thuật học sâu: Mô hình dự đoán và dự báo thời tiết có thể được cải thiện bằng cách áp dụng các kỹ thuật và mô hình học sâu như mạng nơ-ron hồi quy (RNN), mạng nơ-ron tích chập (CNN) hoặc mạng nơ-ron biến đổi (Transformer). Các mô hình học sâu này có khả năng học các đặc trưng phức tạp và tương quan phi tuyến, giúp cải thiện khả năng dự đoán và dự báo thời tiết.

Nâng cao khả năng dự báo thời gian: Mô hình có thể được nâng cấp để dự đoán thời tiết không chỉ trong tương lai gần, mà còn trong tương lai xa hơn. Bằng cách sử dụng quá khứ và dữ liệu hiện tại, cùng với các phương pháp thống kê và xu hướng dự báo, mô hình có thể cung cấp dự báo thời gian dài hơn, giúp người dùng lập kế hoạch cho những sự kiện xa trong tương lai.

# TÀI LIỆU THAM KHẢO

**1. Học máy (Machine learning) là gì? Ứng dụng thực tiễn**, trang TTĐTTH của công ty VietNewsCorp, link: <https://vietnambiz.vn/may-hoc-machine-learning-la-gi-ung-dung-thuc-tien-20190923225908014.htm>

**2. Lavender (2018), Machine Learning – Thử làm nhà thiên văn dự báo thời tiết**, link: <https://viblo.asia/p/machine-learning-thu-lam-nha-thien-van-du-bao-thoi-tiet-djeZ1xYmKWz>

**3. Link data Kaggle:** <https://www.kaggle.com/datasets/ananthr1/weather-prediction>

**4. R. J. Hodges, Atmospheric Science: An Introductory Survey, Cambridge University Pres 2007**, link: <https://www.gnssx.ac.cn/docs/Atmospheric%20Science%20An%20Introductory%20Survey%20(John%20M.%20Wallace,%20Peter%20V.%20Hobbs)%20(z-lib.org).pdf>