

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

=====***=====



BÁO CÁO BTL THUỘC HỌC PHẦN:
TRÍ TUỆ NHÂN TẠO

TÌM HIỂU THUẬT TOÁN NAÏVE BAYES
VÀ ỨNG DỤNG DỰ BÁO THỜI TIẾT

GVHD: Th.S Mai Thanh Hồng

Nhóm: 7

Lớp: 20242IT6094003

Thành viên: Đinh Xuân Hoàng - 2023600793

Đặng Khánh Huyền - 2023600739

Nguyễn Đức Hùng - 2023601298

Lê Nguyễn Hoàng Sơn - 2023602548

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

=====***=====



BÁO CÁO BTL THUỘC HỌC PHẦN:
TRÍ TUỆ NHÂN TẠO

TÌM HIỂU THUẬT TOÁN NAÏVE BAYES
VÀ ỨNG DỤNG DỰ BÁO THỜI TIẾT

GVHD: Th.S Mai Thanh Hồng

Nhóm: 7

Lớp: 20242IT6094003

Thành viên: Đinh Xuân Hoàng - 2023600793

Đặng Khánh Huyền - 2023600739

Nguyễn Đức Hùng - 2023601298

Lê Nguyễn Hoàng Sơn - 2023602548

Hà nội, Năm 2025

BÁO CÁO HỌC TẬP NHÓM

Tên lớp: 20242IT6094003

Khoá: 18

Tên nhóm: 7

Họ và tên thành viên trong nhóm:

(1)Họ và tên SV: Đặng Khánh Huyền

Mã SV: 2023600739

(2)Họ và tên SV: Đinh Xuân Hoàng

Mã SV: 2023600793

(3)Họ và tên SV: Nguyễn Đức Hùng

Mã SV: 2023601298

(4)Họ và tên SV: Lê Nguyễn Hoàng Sơn

Mã SV: 2023602548

Tên chủ đề: Tìm hiểu thuật toán Naïve bayes và ứng dụng dự báo thời tiết

Tuần	Người thực hiện	Nội dung công việc	Kết quả đạt được	Kiến nghị với giảng viên hướng dẫn (Nêu những khó khăn, hỗ trợ từ phía giảng viên, ... nếu cần)
1	Cả nhóm	Viết lời mở đầu, lời cảm ơn, xây dựng mục lục và các mục cần có trong mục lục	Các thành viên hoàn thành đúng thời hạn và đầy đủ nội dung	
2	Đinh Xuân Hoàng	Phân chia công việc chương I cho từng thành viên và tổng hợp.	Các thành viên hoàn thành đúng thời hạn và đầy đủ nội dung	

3	Đặng Khánh Huyền	Phân chia công việc phần 2.1 cho các thành viên và tổng hợp, viết báo cáo tuần 3	Các thành viên hoàn thành đúng thời hạn và đầy đủ nội dung	
4	Lê Nguyễn Hồng Sơn	Phân chia công việc phần 2.2 cho các thành viên và chỉnh sửa, tổng hợp, viết báo cáo tuần 4	Các thành viên hoàn thành đúng thời hạn và đầy đủ nội dung	
5	Nguyễn Đức Hùng	Thảo luận xây dựng code cho bài toán và viết kết luận	Các thành viên hoàn thành đúng thời hạn và đầy đủ nội dung	
6	Đặng Khánh Huyền	Tổng hợp báo cáo, làm phần 2.3 và chỉnh sửa báo cáo	Các thành viên hoàn thành đúng thời hạn và đầy đủ nội dung	

Ngày tháng năm 2025
XÁC NHẬN CỦA GIẢNG VIÊN
 (Ký)

Mai Thanh Hồng

PHIẾU HỌC TẬP NHÓM

I. Thông tin chung:

1. Tên lớp: 20242IT6094003 Khoá: 18
2. Tên nhóm: 7
3. Họ và tên thành viên trong nhóm:

(1) Họ và tên SV: Đặng Khánh Huyền	Mã SV: 2023600739
(2) Họ và tên SV: Đinh Xuân Hoàng	Mã SV: 2023600793
(3) Họ và tên SV: Nguyễn Đức Hùng	Mã SV: 2023601298
(4) Họ và tên SV: Lê Nguyễn Hoàng Sơn	Mã SV: 2023602548

II. Nội dung học tập:

1. Tên chủ đề: Ứng dụng thuật toán Naïve Bayes trong dự báo thời tiết
2. Hoạt động của sinh viên:
 - Hoạt động 1: Đề xuất chủ đề nghiên cứu
 - + Nội dung:
 - Viết đề xuất lựa chọn chủ đề nghiên cứu và xin ý kiến người hướng dẫn về chủ đề nghiên cứu
 - Lập biên bản họp và làm việc nhóm
 - Đặt ra các quy tắc làm việc nhóm:
 - + Mục tiêu/chuẩn đầu ra: Lập biên bản họp và làm việc nhóm
 - Hoạt động 2: Báo cáo tiến độ lần 1
 - Mục tiêu/chuẩn đầu ra:
 - Viết được nội dung phần mở đầu, cảm ơn và chương 1
 - Chương 1: Giới thiệu về thuật toán Naïve Bayes
 - Giới thiệu tổng quan về định lý Bayes và thuật toán Naïve Bayes, các loại mô hình Naïve Bayes, ưu nhược điểm của thuật toán và ứng dụng thực tiễn .
 - Hoạt động 3: Báo cáo tiến độ lần 2
 - Mục tiêu/chuẩn đầu ra:

- Viết được nội dung Chương 2: Xây dựng chương trình.
 - Tiến hành mô tả bài toán đã được đưa ra và sử dụng ngôn ngữ lập trình Python để ứng dụng thuật toán Naïve Bayes trong dự báo thời tiết.
- Hoạt động 4: Nộp cuốn báo cáo thí nghiệm/thực nghiệm và chương trình code

Mục tiêu/chuẩn đầu ra: Hoàn thành và nộp sản phẩm nghiên cứu

3. Sản phẩm nghiên cứu: Quyển báo cáo thí nghiệm/thực nghiệm + Chương trình code

III. Nhiệm vụ học tập:

1. Hoàn thành báo cáo thí nghiệm/thực nghiệm theo đúng thời gian quy định (từ ngày 09/04/2025 đến ngày 25/05/2025)
2. Báo cáo sản phẩm nghiên cứu theo chủ đề được giao trước giảng viên và những sinh viên khác

IV. Học liệu thực hiện Bài tập lớn:

1. Tài liệu học tập:

[1] Nguyễn Phương Nga, Trần Hùng Cường, *Giáo trình Trí tuệ nhân tạo*, NXB Thống kê, 2021.

[2] Nguyễn Thanh Thủy, *Trí tuệ nhân tạo*, NXB Thống kê, 1999.

[3] Nguyễn Đình Thúc, *Giáo trình Trí tuệ nhân tạo: Mạng nơ-ron phương pháp và ứng dụng*, NXB Giáo dục, 2000.

2. Phương tiện, nguyên liệu thực hiện Bài tập lớn: Máy tính cá nhân, máy chiếu, mạng internet.

KẾ HOẠCH LÀM VIỆC NHÓM

1. Tên lớp: 20242IT6094003 Khóa: 18

2. Nhóm: 7

3. Ngày bắt đầu: 07/04/2025

4. Ngày kết thúc: 26/05/2025

5. Thành viên nhóm:

(1)Ho và tên SV: Đặng Khánh Huyền Mã SV: 2023600739

(2) Họ và tên SV: Đinh Xuân Hoàng Mã SV: 2023600793

(3)Ho và tên SV: Nguyễn Đức Hùng Mã SV: 2023601298

(4) **Họ và tên SV:** Lê Nguyễn Hoàng Sơn **Mã SV:** 2023602548

#	Công việc	Ngày bắt đầu dự kiến	Ngày bắt đầu thực tế	Ngày kết thúc dự kiến	Ngày kết thúc thực tế	Trạng thái	Người thực hiện	Ghi chú
1	Viết lời mở đầu	07/04/2025	07/04/2025	13/04/2025	13/04/2025	Done	Nguyễn Đức Hùng	
2	Viết lời cảm ơn	07/04/2025	07/04/2025	13/04/2025	13/04/2025	Done	Lê Nguyễn Hoàng Sơn	
3	Xây dựng mục lục	07/04/2025	07/04/2025	13/04/2025	13/04/2025	Done	Đinh Xuân Hoàng, Đặng Khánh Huyền	

4	Cơ sở lý thuyết	14/04/2025	14/04/2025	20/04/2025	20/04/2025	Done	Lê Nguyễn Hoàng Sơn	
5	Mô hình Gaussian Naive Bayes	14/04/2025	14/04/2025	20/04/2025	20/04/2025	Done	Nguyễn Đức Hùng	
6	Mô hình Multinomial Naive Bayes	14/04/2025	14/04/2025	20/04/2025	20/04/2025	Done	Đinh Xuân Hoàng	
7	Mô hình Bernoulli Naive Bayes	14/04/2025	14/04/2025	20/04/2025	20/04/2025	Done	Đặng Khánh Huyền	
8	Mô tả bài toán	21/04/2025	21/04/2025	27/04/2025	27/04/2025	Done	Đinh Xuân Hoàng, Đặng Khánh Huyền	
9	Những dữ liệu cần dùng để thực hiện	21/04/2025	21/04/2025	27/04/2025	27/04/2025	Done	Nguyễn Đức Hùng	
10	Các đặc trưng đầu vào và đầu	21/04/2025	21/04/2025	27/04/2025	27/04/2025	Done	Lê Nguyễn Hoàng Sơn	

	ra của bài toán							
11	Tiền xử lý dữ liệu	28/04/2025	28/04/2025	03/05/2025	03/05/2025	Done	Đặng Khánh Huyền	
12	Mô hình hoá bài toán với Naive Bayes	28/04/2025	28/04/2025	03/05/2025	03/05/2025	Done	Nguyễn Đức Hùng	
13	Cài đặt mô hình	03/05/2025	03/05/2025	20/05/2025	20/05/2025	Done	Sơn, Huyền, Hùng, Hoàng	
14	Kết quả bài toán	20/05/2025	20/05/2025	25/05/2025	25/05/2025	Done	Đặng Khánh Huyền	
15	Viết kết luận	20/05/2025	20/05/2025	25/05/2025	25/05/2025	Done	Nguyễn Đức Hùng	

BIÊN BẢN HỌP, LÀM VIỆC NHÓM TUẦN 1					
Nhóm: 7					
Thời gian - time: 07/04/2025					
Địa điểm - location: Google meet					
Người chủ trì cuộc họp - chair meeting: Đặng Khánh Huyền					
Thành viên tham dự - Participants : Đặng Khánh Huyền, Đinh Xuân Hoàng, Nguyễn Đức Hùng, Lê Nguyễn Hoàng Sơn					
#	Thành viên	Đúng giờ - On time	Muộn - Late	Vắng - Absent	Ghi chú
1	Đặng Khánh Huyền	V			Done
2	Đinh Xuân Hoàng	V			Done
3	Nguyễn Đức Hùng	V			Done
4	Lê Nguyễn Hoàng Sơn	V			Done
Chương trình họp - Meeting agenda					
#	Mục nội dung - Item	Người trình bày - Owner(s)	Thời gian - Times	Ghi chú , trao đổi - Note	
1	Quản lý nhóm	Cả nhóm	15 phút	-Thiết lập nội quy cho cả nhóm: +Thời gian họp lần tiếp theo được xác định vào lúc cuối của lần trước đó +Các thành viên có trách nhiệm vào đúng giờ mỗi lần họp, hoàn thành các nhiệm được giao nếu không thì bị kiến nghị với giáo viên +Các thành viên tôn trọng lẫn nhau, không phán xét, chỉ trích	

2	Thiết lập kênh giao tiếp, lưu trữ	Đặng Khánh Huyền	5 phút	- Kênh giao tiếp của cả nhóm thông qua: Zalo - Kênh lưu trữ: Google drive
3	Xác định mục tiêu làm việc nhóm	Cả nhóm	5 phút	- Hoàn thành bài tập đúng hạn, thuần thục phương pháp, kỹ năng phân tích thiết kế phần mềm
4	Xác định đề tài	Cả nhóm	30 phút	Thống nhất đề tài làm BTL
5	Phân chia công việc cho từng thành viên	Đặng Khánh Huyền	5 phút	- Đinh Xuân Hoàng, Đặng Khánh Huyền: Tìm hiểu lý thuyết nền tảng liên quan đến đề tài - Nguyễn Đức Hùng, Lê Nguyễn Hoàng Sơn: Thu thập tài liệu tham khảo

Vấn đề & Giải pháp - Issues/problems & Solutions

#	Vấn đề - Issues / problems	Các giải pháp đề xuất - Suggested solutions	Giải pháp được chọn - Selected solution	Ghi chú - Notes
1	Xác định đề tài	Tìm hiểu các thuật toán tìm kiếm mù và ứng dụng vào bài toán rót nước; Tìm hiểu thuật toán Naïve Bayes và ứng dụng dự báo thời tiết, Tìm hiểu thuật toán tìm kiếm heuristic và ứng dụng vào bài toán tìm đường đi ngắn nhất	Đề tài: Tìm hiểu thuật toán Naïve Bayes và ứng dụng dự báo thời tiết	

2	Định lý Bayes và thuật toán Naïve Bayes	Tham khảo trên mạng, tự đọc	Tìm kiếm	
3	Các mô hình và ưu nhược, ứng dụng về các mô hình	Tham khảo trên mạng, tự đọc	Tìm hiểu tài liệu trên mạng	

Kế hoạch hành động - Action plan

#	Hành động - Action	Thời gian - Deadline	Người thực hiện - Owner(s)	Ghi chú - Notes
1	Tìm hiểu cơ sở lý thuyết	13/04/2025	Lê Nguyễn Hoàng Sơn	Toàn bộ phần 1.1
2			Đặng Khánh Huyền	Mô hình Bernoulli Naïve Bayes + ưu nhược điểm và ứng dụng
			Đinh Xuân Hoàng	Mô hình Multinomial Naïve Bayes + ưu nhược điểm và ứng dụng
			Nguyễn Đức Hùng	Mô hình Gaussian Naïve Bayes + ưu nhược điểm và ứng dụng
3	Viết Lời mở đầu và Mục lục cho đề tài	13/04/2025	Đinh Xuân Hoàng	Thu thập công việc hoàn thành của các thành viên

Đóng góp nhóm - Team contribution

#	Thành viên - Members	Ý tưởng , giải pháp - Idea(s)	Hỗ trợ người khác - Support other(s)	Hoạt động xây dựng nhóm - Team bulding activities	Ghi chú - Notes
1	Đặng Khánh Huyền	3	2	2	

2	Đinh Xuân Hoàng	2	3	2	
3	Nguyễn Đức Hùng	2	2	2	
4	Lê Nguyễn Hoàng Sơn	2	2	2	
Kết quả đánh giá phản hồi của nhóm - Team feedback					
#	Số phiếu 4	Số phiếu 3	Số phiếu 2	Số phiếu 1	
1	4	0	0	0	

BIÊN BẢN HỌP, LÀM VIỆC NHÓM TUẦN 2					
Nhóm: 7					
Thời gian - time: 14/04/2025					
Địa điểm - location: Google meet					
Người chủ trì cuộc họp - chair meeting: Đặng Khánh Huyền					
Thành viên tham dự - Participants : Đặng Khánh Huyền, Đinh Xuân Hoàng, Nguyễn Đức Hùng, Lê Nguyễn Hoàng Sơn					
#	Thành viên	Đúng giờ - On time	Muộn - Late	Vắng - Absent	Ghi chú
1	Đặng Khánh Huyền	V			Done
2	Đinh Xuân Hoàng	V			Done
3	Nguyễn Đức Hùng	V			Done
4	Lê Nguyễn Hoàng Sơn	V			Done
Chương trình họp - Meeting agenda					
#	Mục nội dung - Item	Người trình bày - Owner(s)	Thời gian - Times	Ghi chú , trao đổi - Note	
1	Cơ sở lý thuyết: Định lý Bayes và thuật toán Naive Bayes	Lê Nguyễn Hoàng Sơn	15 phút		
2	Mô hình Bernoulli Naïve Bayes, ưu nhược điểm và ứng dụng của mô hình trên	Đặng Khánh Huyền	15 phút		
3	Mô hình Multinomial Naïve Bayes, ưu nhược điểm và ứng dụng của mô hình trên	Đinh Xuân Hoàng	15 phút		

4	Mô hình Gaussian Naïve Bayes, ưu nhược điểm và ứng dụng của mô hình trên	Nguyễn Đức Hùng	15 phút	
---	--	-----------------	---------	--

Vấn đề & Giải pháp - Issues / problems & Solutions

#	Vấn đề - Issues / problems	Các giải pháp đề xuất - Suggested solutions	Giải pháp được chọn - Selected solution	Ghi chú - Notes
1	Ưu điểm, nhược điểm và ứng dụng có sự trùng lặp ý	Tham khảo, tìm hiểu chương trình có sẵn, xem các tài liệu trên học kết hợp	Tham khảo, tìm hiểu chương trình có sẵn	
2	Biểu đồ tương đối dài và còn xung đột ý kiến	Nghe ý kiến của các thành viên, sau đó chọn phương án tốt nhất	Nghe ý kiến của các thành viên, sau đó chọn phương án tốt nhất	

Kế hoạch hành động - Action plan

#	Hành động - Action	Thời gian - Deadline	Người thực hiện - Owner(s)	Ghi chú - Notes
1	Cơ sở lý thuyết: Định lý Bayes và thuật toán Naive Bayes	19/4/2025	Lê Nguyễn Hoàng Sơn	
2	Mô hình Bernoulli Naïve Bayes, ưu nhược điểm và ứng dụng của mô hình trên	19/4/2025	Đặng Khánh Huyền	
3	Mô hình Multinomial Naïve Bayes, ưu nhược điểm và ứng dụng của mô hình trên	19/4/2025	Đinh Xuân Hoàng	

4	Mô hình Gaussian Naïve Bayes, ưu nhược điểm và ứng dụng của mô hình trên	19/4/2025	Nguyễn Đức Hùng		
Đóng góp nhóm - Team contribution					
#	Thành viên - Members	Ý tưởng , giải pháp - Idea(s)	Hộ trợ người khác - Support other(s)	Hoạt động xây dựng nhóm - Team bulding activities	Ghi chú - Notes
1	Đặng Khánh Huyền	2	2	2	
2	Đinh Xuân Hoàng	2	2	2	
3	Nguyễn Đức Hùng	2	3	3	
4	Lê Nguyễn Hoàng Sơn	2	2	2	
Kết quả đánh giá phản hồi của nhóm - Team feedback					
#	Số phiếu 4	Số phiếu 3	Số phiếu 2	Số phiếu 1	
1	4	0	0	0	

BIÊN BẢN HỌP, LÀM VIỆC NHÓM TUẦN 3					
Nhóm: 7					
Thời gian - time: 21/04/2025					
Địa điểm - location: Google meet					
Người chủ trì cuộc họp - chair meeting: Đặng Khánh Huyền					
Thành viên tham dự - Participants : Đặng Khánh Huyền, Đinh Xuân Hoàng, Nguyễn Đức Hùng, Lê Nguyễn Hoàng Sơn					
#	Thành viên	Đúng giờ - On time	Muộn - Late	Vắng - Absent	Ghi chú
1	Đặng Khánh Huyền	V			Done
2	Đinh Xuân Hoàng	V			Done
3	Nguyễn Đức Hùng	V			Done
4	Lê Nguyễn Hoàng Sơn	V			Done
Chương trình họp - Meeting agenda					
#	Mục nội dung - Item	Người trình bày - Owner(s)	Thời gian - Times	Ghi chú , trao đổi - Note	
1	Xây dựng các yêu cầu cho các phần	Đặng Khánh Huyền	15 phút		
2	Tổng quan về bài toán dự báo thời tiết	Đinh Xuân Hoàng	15 phút		
3	Tìm kiếm các tập dữ liệu sử dụng	Nguyễn Đức Hùng	15 phút		
4	Các đặc trưng đầu vào đầu ra	Lê Nguyễn Hoàng Sơn	15 phút		

Vấn đề & Giải pháp - Issues / problems & Solutions					
#	Vấn đề - Issues / problems	Các giải pháp đề xuất - Suggested solutions		Giải pháp được chọn - Selected solution	Ghi chú - Notes
1	Bài toán tương đối nặng và khó	Tham khảo, tìm hiểu chương trình có sẵn, xem các tài liệu trên học kết hợp		Tham khảo, tìm hiểu chương trình có sẵn	
Kế hoạch hành động - Action plan					
#	Hành động - Action	Thời gian - Deadline	Người thực hiện - Owner(s)	Ghi chú - Notes	
1	Xây dựng các yêu cầu cho các phần	27/04/2025	Đặng Khánh Huyền		
2	Tổng quan về bài toán dự báo thời tiết	27/04/2025	Đinh Xuân Hoàng		
3	Tìm kiếm các tập dữ liệu sử dụng	27/04/2025	Nguyễn Đức Hùng		
4	Các đặc trưng đầu vào đầu ra	27/04/2025	Lê Nguyễn Hoàng Sơn		
Đóng góp nhóm - Team contribution					
#	Thành viên - Members	Ý tưởng , giải pháp - Idea(s)	Hỗ trợ người khác - Support other(s)	Hoạt động xây dựng nhóm - Team bulding activities	Ghi chú - Notes
1	Đặng Khánh Huyền	2	2	2	
2	Đinh Xuân Hoàng	2	3	3	
3	Nguyễn Đức Hùng	2	2	2	

4	Lê Nguyễn Hoàng Sơn	2	2	2	
Kết quả đánh giá phản hồi của nhóm - Team feedback					
#	Số phiếu 4	Số phiếu 3	Số phiếu 2	Số phiếu 1	
1	4	0	0	0	

BIÊN BẢN HỌP, LÀM VIỆC NHÓM TUẦN 4					
Nhóm: 7					
Thời gian - time: 28/04/2025					
Địa điểm - location: Google meet					
Người chủ trì cuộc họp - chair meeting: Đinh Xuân Hoàng					
Thành viên tham dự - Participants : Đinh Xuân Hoàng, Đặng Khánh Huyền, Nguyễn Đức Hùng, Lê Nguyễn Hoàng Sơn					
#	Thành viên	Đúng giờ - On time	Muộn - Late	Vắng - Absent	Ghi chú
1	Đặng Khánh Huyền	V			Done
2	Lê Nguyễn Hoàng Sơn	V			Done
3	Đinh Xuân Hoàng	V			Done
4	Nguyễn Đức Hùng	V			Done
Chương trình họp - Meeting agenda					
#	Mục nội dung - Item	Người trình bày - Owner(s)	Thời gian - Times	Ghi chú , trao đổi - Note	
1	Tiền xử lý dữ liệu cho bài toán	Đặng Khánh Huyền	15 phút		
2	Mô hình hoá dữ liệu với Naive Bayes	Nguyễn Đức Hùng	15 phút		
3	Cài đặt mô hình Naive Bayes	Lê Nguyễn Hoàng Sơn	15 phút		
4	Tổng hợp nội dung, chỉnh sửa, viết báo cáo	Đinh Xuân Hoàng	15 phút		
Vấn đề & Giải pháp - Issues / problems & Solutions					
#	Vấn đề - Issues / problems	Các giải pháp đề xuất - Suggested solutions		Giải pháp được chọn - Selected solution	Ghi chú - Notes

1	Cấu trúc về mô hình chưa biết cấu trúc như thế nào về một bài toán dự báo	Tham khảo qua mạng , bài giảng của giáo viên, hệ thống kết hợp	Tham khảo cấu trúc trên mạng		
Kế hoạch hành động - Action plan					
#	Hành động - Action	Thời gian - Deadline	Người thực hiện - Owner(s)	Ghi chú - Notes	
1	Tiền xử lý dữ liệu cho bài toán	3/5/2025	Đặng Khánh Huyền		
2	Mô hình hoá dữ liệu với Naive Bayes	3/5/2025	Nguyễn Đức Hùng		
3	Cài đặt mô hình Naive Bayes	3/5/2025	Lê Nguyễn Hoàng Sơn		
4	Tổng hợp nội dung, chỉnh sửa, viết báo cáo	3/5/2025	Đinh Xuân Hoàng		
Đóng góp nhóm - Team contribution					
#	Thành viên - Members	Ý tưởng , giải pháp - Idea(s)	Hộ trợ người khác - Support other(s)	Hoạt động xây dựng nhóm - Team bulding activities	Ghi chú - Notes
1	Đặng Khánh Huyền	2	2	2	
2	Nguyễn Đức Hùng	2	2	2	
3	Lê Nguyễn Hoàng Sơn	2	2	2	
4	Đinh Xuân Hoàng	2	2	2	
Kết quả đánh giá phản hồi của nhóm - Team feedback					
#	Số phiếu 4	Số phiếu 3	Số phiếu 2	Số phiếu 1	
1	4	0	0	0	

BIÊN BẢN HỌP, LÀM VIỆC NHÓM TUẦN 5					
Nhóm: 7					
Thời gian - time: 05/05/2025					
Địa điểm - location: Google meet					
Người chủ trì cuộc họp - chair meeting: Đinh Xuân Hoàng					
Thành viên tham dự - Participants : Đinh Xuân Hoàng, Đặng Khánh Huyền, Nguyễn Đức Hùng, Lê Nguyễn Hoàng Sơn					
#	Thành viên	Đúng giờ - On time	Muộn - Late	Vắng - Absent	Ghi chú
1	Đặng Khánh Huyền	V			Done
2	Lê Nguyễn Hoàng Sơn	V			Done
3	Đinh Xuân Hoàng	V			Done
4	Nguyễn Đức Hùng	V			Done
Chương trình họp - Meeting agenda					
#	Mục nội dung - Item	Người trình bày - Owner(s)	Thời gian - Times	Ghi chú , trao đổi - Note	
1	Tìm hiểu về các dữ liệu mô phỏng thời tiết và support Hoàng	Đặng Khánh Huyền	15 phút		
2	Tìm hiểu về cách phân chia dữ liệu và huấn	Nguyễn Đức Hùng	15 phút		

	luyện mô hình			
3	Tìm hiểu về hàm dự đoán kết quả và đánh giá	Lê Nguyễn Hoàng Sơn	15 phút	
4	Tìm hiểu về code phân dự báo thời tiết và thử nghiệm	Đinh Xuân Hoàng	15 phút	

Vấn đề & Giải pháp - Issues / problems & Solutions

#	Vấn đề - Issues / problems	Các giải pháp đề xuất - Suggested solutions	Giải pháp được chọn - Selected solution	Ghi chú - Notes
1	Chưa có nhiều kiến thức về ngôn ngữ python cũng như các hàm sử dụng trong python do vậy còn nhiều khó khăn trong việc thực hiện xây dựng code demo bài toán	Tham khảo qua mạng , bài giảng của giáo viên, hệ thống kết hợp	Tham khảo cấu trúc trên mạng	

Kế hoạch hành động - Action plan

#	Hành động - Action	Thời gian - Deadline	Người thực hiện - Owner(s)	Ghi chú - Notes	
1	Tìm hiểu về các dữ liệu mô phỏng thời tiết và support Hoàng	19/5/2025	Đặng Khánh Huyền		
2	Tìm hiểu về cách phân chia dữ liệu và huấn luyện mô hình	19/5/2025	Nguyễn Đức Hùng		
3	Tìm hiểu về hàm dự đoán kết quả và đánh giá	19/5/2025	Lê Nguyễn Hoàng Sơn		
4	Tìm hiểu về code phần dự báo thời tiết và thử nghiệm	19/5/2025	Đinh Xuân Hoàng		
Đóng góp nhóm - Team contribution					
#	Thành viên - Members	Ý tưởng , giải pháp - Idea(s)	Hỗ trợ người khác - Support other(s)	Hoạt động xây dựng nhóm - Team bulding activities	Ghi chú - Notes
1	Đặng Khánh Huyền	2	2	2	

2	Nguyễn Đức Hùng	2	2	2	
3	Lê Nguyễn Hoàng Sơn	2	2	3	
4	Đinh Xuân Hoàng	2	2	2	
Kết quả đánh giá phản hồi của nhóm - Team feedback					
#	Số phiếu 4	Số phiếu 3	Số phiếu 2	Số phiếu 1	
1	4	0	0	0	

MỤC LỤC

DANH MỤC HÌNH ẢNH	3
LỜI NÓI ĐẦU	4
MỞ ĐẦU	5
1. Lý do chọn đề tài	5
2. Mục đích nghiên cứu	5
3. Đối tượng nghiên cứu	5
4. Phạm vi nghiên cứu	5
5. Kết quả mong muốn đạt được của đề tài	6
6. Cấu trúc của báo cáo	6
CHƯƠNG I: GIỚI THIỆU VỀ THUẬT TOÁN NAÏVE BAYES	7
1.1. Định lý Bayes và thuật toán Naïve Bayes	7
1.1.1. Định lý bayes	7
1.1.2. Giả định Naive (độc lập có điều kiện)	8
1.1.3. Công thức tổng quát	9
1.1.4. Độ phức tạp của thuật toán Naïve Bayes	9
1.2. Các loại mô hình	9
1.2.1. Gaussian Naïve Bayes	9
1.2.2. Multinomial Naïve Bayes	10
1.2.3. Bernoulli Naïve Bayes	11
1.3. Ưu điểm, nhược điểm và ứng dụng thực tiễn của thuật toán NAÏVE BAYES	12
1.3.1. Ưu điểm	12
1.3.2. Nhược điểm	13
1.3.3. Ứng dụng thực tiễn	14
CHƯƠNG II: ỨNG DỤNG NAÏVE BAYES TRONG DỰ BÁO THỜI TIẾT	20
2.1. BÀI TOÁN DỰ BÁO THỜI TIẾT	20
2.1.1. Mô tả bài toán:	20
2.1.2. Tập dữ liệu sử dụng	22
2.1.3. Các đặc trưng đầu vào và đầu ra	22
2.2. Phương pháp giải quyết bài toán	25

2.2.1 Tiền xử lý dữ liệu	25
2.2.2. Mô hình hóa với Naïve Bayes	29
2.2.3. Cài đặt mô hình.....	31
2.3 Kết quả	38
KẾT LUẬN	40
TÀI LIỆU THAM KHẢO.....	41

DANH MỤC HÌNH ẢNH

Hình 1.1: Ảnh minh hoạ spam.....	15
Hình 1.2: Ảnh minh hoạ Naïve Bayes.....	16
Hình 1.3: Ví dụ về dự đoán	17
Hình 1.4: Ví dụ về hệ thống đề xuất.....	17
Hình 1.5: Ví dụ về phân loại tin tức	18
Hình 1.6: Ví dụ về nhận diện gian lận.....	19
Hình 2.1: Sơ đồ minh hoạ bài toán dự báo thời tiết	31
Hình 2.2: Ví dụ chạy chương trình.....	38
Hình 2.3: Ví dụ về trời không mưa	39
Hình 2.4: Ví dụ về trời mưa.....	39

LỜI NÓI ĐẦU

Trong bối cảnh công nghệ thông tin ngày càng phát triển, việc áp dụng các thuật toán học máy vào thực tiễn đang ngày càng đóng vai trò then chốt trong nhiều lĩnh vực khác nhau. Báo cáo này được thực hiện với mục tiêu xây dựng một ứng dụng dự báo thời tiết dựa trên thuật toán Naïve Bayes – một trong những thuật toán phân loại cơ bản nhưng hiệu quả của học máy. Đề tài không chỉ hướng tới việc phát triển kỹ năng lập trình và tư duy phân tích dữ liệu của nhóm mà còn mong muốn tạo ra một công cụ học tập trực quan cho những người quan tâm đến lĩnh vực dự báo và phân tích dữ liệu.

Chúng em xin gửi lời cảm ơn chân thành đến cô Mai Thanh Hồng – giảng viên hướng dẫn, người đã tận tình chỉ bảo và đóng góp những ý kiến quý báu trong suốt quá trình thực hiện báo cáo. Đồng thời, chúng em cũng xin chân thành cảm ơn nhà trường và khoa chuyên môn đã tạo điều kiện để nhóm có cơ hội thực hiện đề tài này.

Mặc dù đã cố gắng hết sức, nhưng chắc chắn không thể tránh khỏi những thiếu sót do kiến thức và trình độ chuyên môn còn hạn chế trong quá trình thực hiện nghiên cứu. Vì vậy, chúng em rất mong nhận được ý kiến đóng góp, bổ sung quý báu từ phía thầy cô để đề tài này của chúng em được hoàn thiện hơn trong tương lai.

Chúng em xin chân thành cảm ơn!

Nhóm thực hiện đề tài

Nhóm 7

MỞ ĐẦU

1. Lý do chọn đề tài

Trong thời đại công nghệ số phát triển mạnh mẽ, việc dự đoán và phân tích dữ liệu thời tiết đóng vai trò quan trọng trong nhiều lĩnh vực như nông nghiệp, giao thông, du lịch và đời sống hằng ngày. Thuật toán Naïve Bayes là một trong những thuật toán phân loại đơn giản nhưng hiệu quả, thường được ứng dụng rộng rãi trong các bài toán dự báo với độ chính xác cao. Việc xây dựng một mô hình dự báo thời tiết sử dụng thuật toán Naïve Bayes không chỉ giúp củng cố kiến thức về học máy mà còn tạo ra một công cụ thực tiễn hỗ trợ việc dự đoán thời tiết trong thực tế.

2. Mục đích nghiên cứu

Mục đích chính của đề tài là xây dựng một hệ thống dự báo thời tiết dựa trên việc phân tích các dữ liệu khí tượng bằng thuật toán Naïve Bayes. Đề tài nhằm giúp nhóm nâng cao kỹ năng lập trình, tư duy xử lý dữ liệu, và tiếp cận thực tế hơn với các ứng dụng của học máy trong cuộc sống hằng ngày.

3. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài là thuật toán Naïve Bayes và các phương pháp ứng dụng thuật toán này trong việc phân loại và dự báo dữ liệu thời tiết dựa trên các đặc trưng đầu vào.

4. Phạm vi nghiên cứu

Đề tài tập trung vào:

- Xây dựng mô hình dự báo thời tiết đơn giản trên dữ liệu thời tiết có sẵn.
- Áp dụng thuật toán Naïve Bayes để dự đoán các yếu tố như tình trạng thời tiết (nắng, mưa, nhiều mây,...) dựa trên các thông số khí tượng như nhiệt độ, độ ẩm, áp suất.
- Thiết kế giao diện trực quan, dễ sử dụng để hiển thị kết quả dự báo.

Đề tài không đi sâu vào các mô hình học sâu (Deep Learning) hay các hệ thống dự báo thời tiết quy mô lớn.

5. Kết quả mong muốn đạt được của đề tài

- Xây dựng thành công mô hình dự báo thời tiết sử dụng thuật toán Naïve Bayes.
- Tạo ra một công cụ đơn giản hỗ trợ người dùng hoặc sinh viên có thể hiểu rõ hơn về cách hoạt động của thuật toán phân loại trong thực tiễn.
- Hoàn thiện báo cáo nghiên cứu chi tiết, trình bày rõ quy trình thực hiện và kết quả đạt được.

6. Cấu trúc của báo cáo

Báo cáo bài tập lớn gồm 2 chương (ngoài phần mở đầu):

- **Chương 1:** Cơ sở lý thuyết
- **Chương 2:** Thực hiện bài toán

CHƯƠNG I: GIỚI THIỆU VỀ THUẬT TOÁN NAÏVE BAYES

1.1. Định lý Bayes và thuật toán Naïve Bayes

1.1.1. Định lý bayes

Định lý Bayes là một trong những định lý quan trọng trong lý thuyết xác suất, giúp tính toán xác suất xảy ra của một sự kiện dựa trên thông tin đã biết về các sự kiện liên quan.

Công thức định lý Bayes được phát biểu như sau:

$$P(h | D) = \frac{P(D | h)P(h)}{P(D)}$$

Trong đó:

- $P(A|B)$: Xác suất xảy ra của sự kiện A khi biết B đã xảy ra (xác suất hậu nghiệm).
- $P(D)$: xác suất trước (tiên nghiệm) của việc quan sát được dữ liệu D.
- $P(D | h)$: xác suất có điều kiện của việc quan sát được dữ liệu D, nếu biết giả thiết phân lớp h là đúng.
- $P(h | D)$: xác suất có điều kiện của giả thiết phân lớp h là đúng, nếu quan sát được dữ liệu D.

$$\begin{aligned} h_{\text{MAP}} &= \underset{h \in H}{\operatorname{argmax}} P(h | D) \\ &= \underset{h \in H}{\operatorname{argmax}} \frac{P(D | h).P(h)}{P(D)} \\ &= \underset{h \in H}{\operatorname{argmax}} P(D | h).P(h) \end{aligned}$$

Với H là một tập các giả thiết phân lớp, hệ thống học sẽ tìm giả thiết có thể xảy ra nhất để $h \in H$ đối với các dữ liệu D quan sát được. Giả thiết h này được gọi là giả thiết có xác suất hậu nghiệm cực đại (Maximum a posteriori – MAP).

Giả thiết có thể nhất $h_{MAP} = h_1$ nếu $P(h_1 | D) \geq P(h_2 | D)$; ngược lại, $h_{MAP} = h_2$. Do $P(D)$ là như nhau đối với cả hai giả thiết h_1 và h_2 , nên có thể bỏ qua đại lượng $P(D)$. Vì vậy, chỉ cần tính hai biểu thức: $P(D | h_1).P(h_1)$ và $P(D | h_2).P(h_2)$, là ta có thể đưa ra quyết định phân lớp.

Định lý Bayes giúp ta dễ dàng cập nhật và tính toán xác suất khi có thêm thông tin mới về dữ kiện liên quan.

1.1.2. Giả định Naive (độc lập có điều kiện)

Trong thực tế, việc tính toán xác suất $P(Z|C)$ với tập hợp các thuộc tính $Z = (z_1, z_2, \dots, z_n)$ là rất phức tạp nếu các thuộc tính này có sự phụ thuộc lẫn nhau.

Thuật toán Naive Bayes đưa ra một giả định quan trọng nhằm đơn giản hóa quá trình tính toán, đó là: các thuộc tính (đặc trưng) là độc lập có điều kiện với nhau khi đã biết lớp (class) C .

Giả định này được phát biểu như sau:

$$P(z_1, z_2, \dots, z_n | c_i) = \prod_{j=1}^n P(z_j | c_i)$$

Trong đó:

- z_1, z_2, \dots, z_n là các thuộc tính (feature) của mẫu dữ liệu.
- C là lớp (class) của dữ liệu.

Nhờ giả định Naive này, việc tính toán xác suất trở nên đơn giản hơn rất nhiều, vì chỉ cần tính từng xác suất thành phần rồi nhân lại với nhau, thay vì phải xét đến mối quan hệ phức tạp giữa các thuộc tính.

1.1.3. Công thức tổng quát

Từ định lý Bayes và giả định Naïve, ta có công thức tính xác suất một mẫu dữ liệu thuộc lớp C như sau:

$$P(C|x_1, x_2, \dots, x_n) = \frac{P(C) \cdot \prod_{i=1}^n P(x_i|C)}{P(x_1, x_2, \dots, x_n)}$$

Trong đó:

- $P(C)$ là xác suất tiên nghiệm của lớp C.
- $P(x_i|C)$ là xác suất xuất hiện của thuộc tính x_i khi biết rằng mẫu dữ liệu thuộc lớp C.

Vì $P(x_1, x_2, x_3, \dots, x_n)$ là hằng số với mọi lớp C nên ta có thể bỏ qua khi so sánh giữa các lớp. Khi đó, công thức phân loại trở thành:

$$P(C|x_1, x_2, \dots, x_n) \propto P(C) \cdot \prod_{i=1}^n P(x_i|C)$$

1.1.4. Độ phức tạp của thuật toán Naïve Bayes

Mẫu dữ liệu sẽ được gán vào lớp C có giá trị $P(C|x_1, x_2, \dots, x_n)$ lớn nhất.

- Thời gian huấn luyện: $O(n \times m)$ (n: số thuộc tính, m: số mẫu dữ liệu)
- Thời gian phân loại: $O(n)$ (vì chỉ cần tính tích xác suất của n thuộc tính)
- Không gian lưu trữ: $O(n \times k)$ (k: số lớp)

1.2. Các loại mô hình

1.2.1. Gaussian Naïve Bayes

Gaussian Naive Bayes là một phiên bản của thuật toán Naive Bayes, được sử dụng trong các bài toán phân loại khi các đặc trưng đầu vào là dữ liệu liên tục, chẳng hạn như chiều cao, cân nặng, nhiệt độ, hoặc các giá trị số khác. Thuật toán này kế thừa

nguyên lý cốt lõi của Naive Bayes, dựa trên định lý Bayes và giả định rằng các đặc trưng là độc lập với nhau (naive assumption). Điểm khác biệt chính của Gaussian Naive Bayes nằm ở việc nó giả định rằng xác suất của các đặc trưng trong mỗi lớp tuân theo phân phối chuẩn (Gaussian distribution).

Nguyên lý hoạt động:

Trong Gaussian Naive Bayes, ta giả định rằng các đặc trưng x_i (với $i=1,2,...,n$) theo phân phối chuẩn với giá trị trung bình μ và độ lệch chuẩn σ . Xác suất có điều kiện $P(x_i|c)$ được tính theo công thức của phân phối chuẩn:

$$P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{x_i - \mu_c}{2\sigma_c^2}\right)$$

Trong đó:

- μ_c là giá trị trung bình của đặc trưng x_i trong lớp c
- σ_c là độ lệch chuẩn tương ứng

1.2.2. Multinomial Naïve Bayes

Mô hình này chủ yếu được sử dụng trong phân loại văn bản mà feature vectors được tính bằng Bags of Words. Lúc này, mỗi văn bản được biểu diễn bởi một vector có độ dài d chính là số từ trong từ điển. Giá trị của thành phần thứ i trong mỗi vector chính là số lần từ thứ i xuất hiện trong văn bản đó.

Khi đó, $p(x_i|y)$ tỉ lệ với tần suất từ thứ i (hay feature thứ i cho trường hợp tổng quát) xuất hiện trong các văn bản của class y . Giá trị này có thể được tính bằng cách:

$$P(x_i|y) = \frac{N_i}{N_c}$$

Trong đó:

- N_i là tổng số lần từ x_i xuất hiện trong văn bản
- N_c là tổng số lần từ của tất cả các từ x_1, x_2, \dots, x_n xuất hiện trong văn bản

Công thức trên có hạn chế là khi từ x_i không xuất hiện lần nào trong văn bản, ta sẽ có $N_i=0$. Điều này làm cho $P(x_i|y)=0$

Để khắc phục vấn đề này, người ta sử dụng kỹ thuật gọi là Laplace Smoothing bằng cách cộng thêm vào cả tử và mẫu để giá trị luôn khác 0

$$P(x_i|y) = \frac{N_i + \alpha}{N_c + d\alpha}$$

Trong đó:

- α thường là số dương, bằng 1
- $d\alpha$ được cộng vào mẫu để đảm bảo $\sum_{i=1}^d P(x_i|y)=1$

1.2.3. Bernoulli Naïve Bayes

1.2.3.1. Định nghĩa

Bernoulli Naïve Bayes là một biến thể của thuật toán Naïve Bayes, được sử dụng trong các bài toán phân loại mà dữ liệu đầu vào là các đặc trưng nhị phân (binary features).

Mỗi đặc trưng x_i trong vector đặc trưng $X=(x_1, x_2, \dots, x_n)$ chỉ nhận hai giá trị:

$x_i=1$: đặc trưng xuất hiện.

$x_i=0$: đặc trưng không xuất hiện.

Khác với Multinomial Naïve Bayes (làm việc với tần suất từ), Bernoulli Naïve Bayes chỉ quan tâm xem một đặc trưng có xuất hiện hay không trong mẫu dữ liệu, nên rất phù hợp cho các tác vụ như lọc spam, xác định sự xuất hiện từ khóa, hoặc nhận diện chủ đề văn bản với vector hóa dữ liệu dạng Bag-of-Words nhị phân.

1.2.3.2. Nguyên lý hoạt động

Mô hình này được áp dụng cho các loại dữ liệu mà mỗi thành phần là một giá trị binary - bằng 0 hoặc 1. Ví dụ: cũng với loại văn bản nhưng thay vì đếm tổng số lần xuất hiện của 1 từ trong văn bản, ta chỉ cần quan tâm từ đó có xuất hiện hay không.

Khi đó, $p(x_i|c)$ được tính bằng:

$$p(x_i|c) = p(i|c)^{x_i}(1 - p(i|c))^{1-x_i}$$

với $p(i|c)$ có thể được hiểu là xác suất từ thứ i xuất hiện trong các văn bản của class c .

1.3. Ưu điểm, nhược điểm và ứng dụng thực tiễn của thuật toán NAÏVE BAYES

1.3.1. Ưu điểm

1. Dễ triển khai:

Thuật toán phân loại Naive Bayes rất đơn giản và dễ triển khai. Nó không yêu cầu nhiều tính toán hay thời gian huấn luyện. Thuật toán này có thể được sử dụng cho cả bài toán phân loại nhị phân và phân loại nhiều lớp.

2. Xử lý tốt dữ liệu thiếu:

Thuật toán này rất hữu ích trong việc xử lý dữ liệu bị thiếu. Khi đo lường độ chính xác, bộ phân loại này chỉ xem xét dữ liệu hiện có và bỏ qua những dữ liệu không có. Nhờ đó, độ chính xác vẫn được duy trì.

3. Nhanh và có khả năng mở rộng:

Naive Bayes có tốc độ xử lý nhanh và dễ mở rộng, có thể làm việc với tập dữ liệu lớn. Nó phù hợp cho việc học nhanh và các tác vụ phân loại thời gian thực, đồng thời dễ dàng được song song hóa để chạy trên nhiều bộ xử lý hoặc cụm máy tính.

4. Dễ hiểu:

Naive Bayes dễ hiểu vì nó cung cấp giải thích chi tiết về cách phân loại được thực hiện. Dựa trên sự xuất hiện hoặc không xuất hiện của từng đặc trưng, thuật toán xác định xác suất của một kết quả cụ thể và gán lớp dựa trên xác suất cao nhất.

5. Hiệu quả trong phân loại văn bản:

Naive Bayes là thuật toán được ưa chuộng trong các tác vụ phân loại văn bản, như phân tích cảm xúc hoặc lọc thư rác. Nó hoạt động hiệu quả do có khả năng xử lý dữ liệu có số chiều cao và dữ liệu dạng phân loại – những yếu tố phổ biến trong xử lý ngôn ngữ tự nhiên.

6. Hoạt động tốt với tập dữ liệu nhỏ:

Naive Bayes cho kết quả tốt ngay cả với các tập dữ liệu nhỏ vì không cần nhiều dữ

liệu huấn luyện để đưa ra dự đoán đáng tin cậy. Điều này làm cho nó trở thành lựa chọn phù hợp cho các ứng dụng có dữ liệu hạn chế, chẳng hạn như phát hiện gian lận hoặc chẩn đoán y khoa.

7. Chống chịu tốt với đặc trưng không liên quan:

Naive Bayes có khả năng chống chịu tốt với các đặc trưng không liên quan trong tập dữ liệu. Nguyên nhân là do thuật toán giả định tất cả các đặc trưng là độc lập với nhau và tính toán xác suất kết quả dựa trên sự xuất hiện hoặc không của từng đặc trưng một cách độc lập.

8. Cần ít dữ liệu huấn luyện hơn:

So với các thuật toán học máy khác như cây quyết định hoặc mạng nơ-ron, Naive Bayes cần ít dữ liệu huấn luyện hơn. Điều này là do số lượng tham số cần ước lượng từ dữ liệu được giảm đi nhờ giả định về tính độc lập của các đặc trưng.

9. Xử lý được cả dữ liệu liên tục và rời rạc:

Naive Bayes là một thuật toán linh hoạt vì có thể xử lý được cả dữ liệu liên tục lẫn dữ liệu rời rạc. Tùy theo loại dữ liệu, thuật toán sử dụng các phân phối xác suất khác nhau, như phân phối Gaussian hoặc phân phối đa thức (multinomial).

1.3.2. Nhược điểm

1. Giả định về tính độc lập:

Thuật toán đưa ra giả định rằng tất cả các đặc trưng là độc lập với nhau, điều này thường không đúng trong các ứng dụng thực tế. Nếu các đặc trưng có mối tương quan với nhau, điều này có thể dẫn đến kết quả phân loại không chính xác.

2. Thiếu linh hoạt:

Vì Naive Bayes là một mô hình tham số (parametric model), nó yêu cầu một tập các tham số cố định phải được học từ dữ liệu huấn luyện. Điều này có thể hạn chế khả năng của nó trong việc xử lý các mối quan hệ phức tạp và phi tuyến tính giữa các đặc trưng.

3. Thiếu hụt dữ liệu:

Để Naive Bayes ước lượng chính xác xác suất có điều kiện của từng đặc trưng, cần phải có đủ dữ liệu huấn luyện. Nếu dữ liệu không đủ, thuật toán có thể hoạt động kém hiệu quả.

4. Nhạy cảm với giá trị ngoại lai:

Naive Bayes rất nhạy cảm với các giá trị ngoại lai hoặc cực đoan trong dữ liệu, điều này có thể ảnh hưởng lớn đến các xác suất được ước lượng và dẫn đến kết quả phân loại sai.

5. Mất cân bằng lớp:

Khi dữ liệu bị mất cân bằng và một lớp có số lượng mẫu nhiều hơn đáng kể so với lớp còn lại, Naive Bayes có thể gặp khó khăn trong việc xử lý. Điều này có thể dẫn đến sự thiên vị về phía lớp chiếm ưu thế và hiệu suất kém đối với lớp thiểu số.

6. Khó nắm bắt tương tác giữa các đặc trưng:

Naive Bayes có thể không phát hiện được các mối tương tác hoặc phụ thuộc giữa các đặc trưng – những yếu tố có thể quan trọng đối với quá trình phân loại – vì nó giả định rằng các đặc trưng là độc lập với nhau.

7. Khả năng xử lý biến liên tục hạn chế:

Mô hình Naive Bayes giả định rằng các đặc trưng là rời rạc hoặc phân loại, do đó không thể xử lý trực tiếp các biến liên tục. Dữ liệu liên tục cần được rời rạc hóa trước khi sử dụng thuật toán, điều này có thể gây mất thông tin và làm giảm hiệu suất.

8. Thiên lệch đối với các đặc trưng có tần suất cao:

Naive Bayes có xu hướng thiên lệch về phía những đặc trưng xuất hiện thường xuyên trong tập dữ liệu huấn luyện. Điều này có thể trở thành vấn đề nếu một số đặc trưng hiếm gặp nhưng quan trọng lại bị bỏ qua.

9. Khó xử lý dữ liệu thiếu:

Naive Bayes gặp khó khăn trong việc xử lý dữ liệu bị thiếu và thường xử lý không hiệu quả. Nếu một đặc trưng bị thiếu giá trị, toàn bộ mẫu có thể phải bị loại bỏ hoặc thay thế giá trị, điều này có thể tạo ra kết quả thiên lệch.

10. Nhạy cảm với lựa chọn xác suất tiên nghiệm:

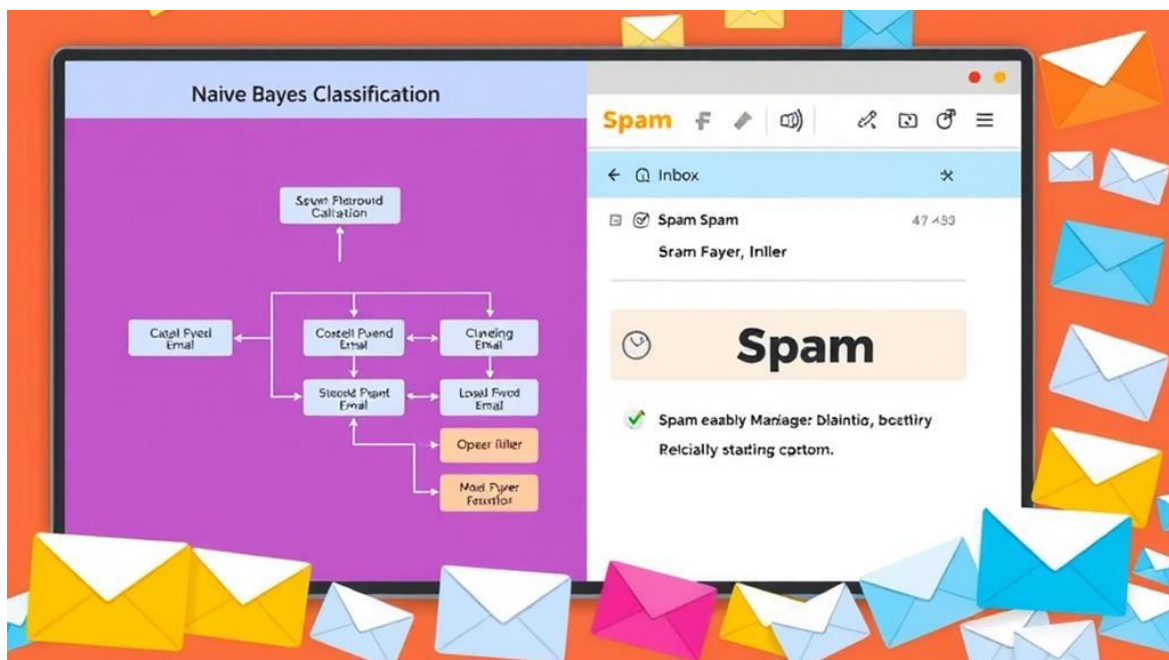
Naive Bayes yêu cầu xác định xác suất tiên nghiệm cho mỗi lớp, và điều này có thể ảnh hưởng đến kết quả phân loại. Hiệu suất của thuật toán có thể bị ảnh hưởng đáng kể bởi việc lựa chọn xác suất tiên nghiệm, vốn có thể mang tính tùy ý.

1.3.3. Ứng dụng thực tiễn

- 1. Lọc spam email:** Phân loại email thành "spam" hoặc "không spam" dựa trên tần suất từ khóa (Multinomial NB)

Quá trình phát hiện spam bắt đầu bằng việc thu thập và phân tích một tập dữ liệu lớn các email hoặc tin nhắn đã được phân loại. Các đặc điểm của những thông điệp này, chẳng hạn như từ ngữ, cụm từ thường gặp, và cấu trúc câu, sẽ được trích xuất và sử dụng để tạo ra mô hình. Naive Bayes sẽ tính toán xác suất của từng đặc điểm liên quan đến việc một thông điệp là spam hay không.

Sau khi mô hình đã được huấn luyện, nó có thể được áp dụng để phân loại các thông điệp mới. Khi một email mới đến, mô hình sẽ xem xét các đặc điểm của email đó và tính toán xác suất tương ứng. Nếu xác suất là cao hơn một ngưỡng nhất định, email đó sẽ được đánh dấu là spam. Phương pháp này không chỉ giúp giảm thiểu số lượng spam mà còn nâng cao hiệu quả trong việc quản lý thông tin cá nhân.

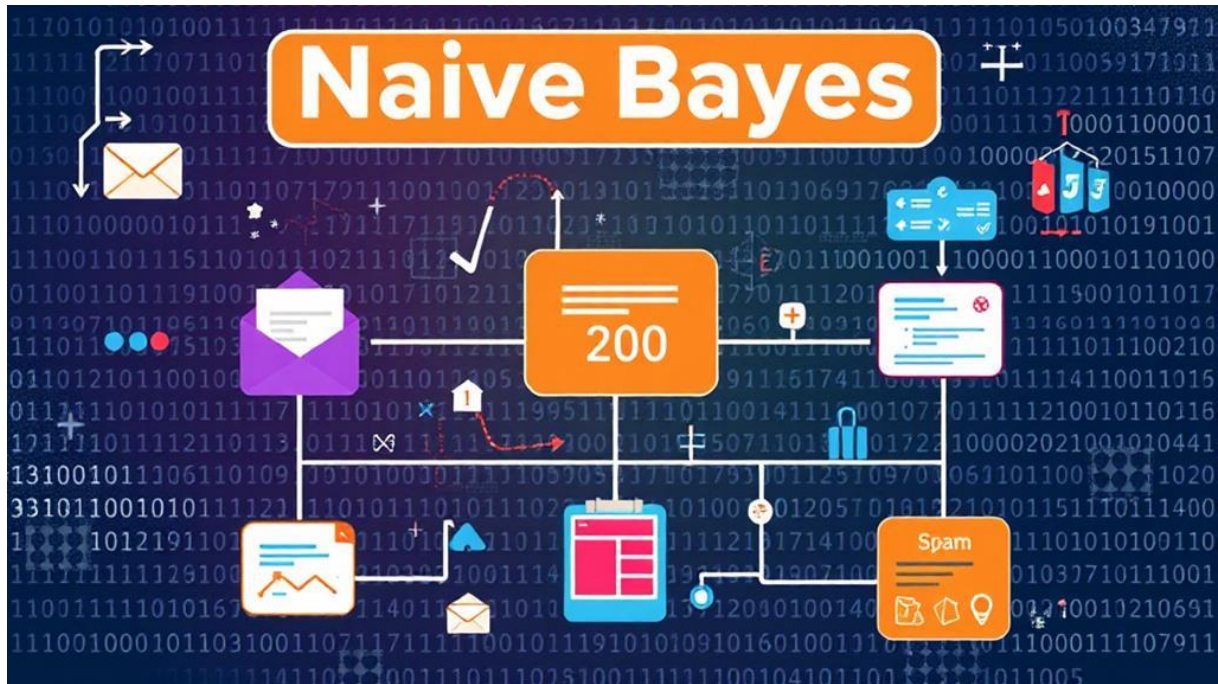


Hình 1.1: Ảnh minh họa spam

2. **Phân loại văn bản:** Được sử dụng trong phân tích cảm xúc, phân loại tài liệu và phân loại chủ đề.

Naive Bayes được áp dụng trong các lĩnh vực như phân loại tin tức, phân tích cảm xúc, và nhận dạng văn bản. Trong phân loại tin tức, nó giúp xác định chủ đề chính của một bài báo dựa trên các từ khóa xuất hiện trong nội dung. Đối với

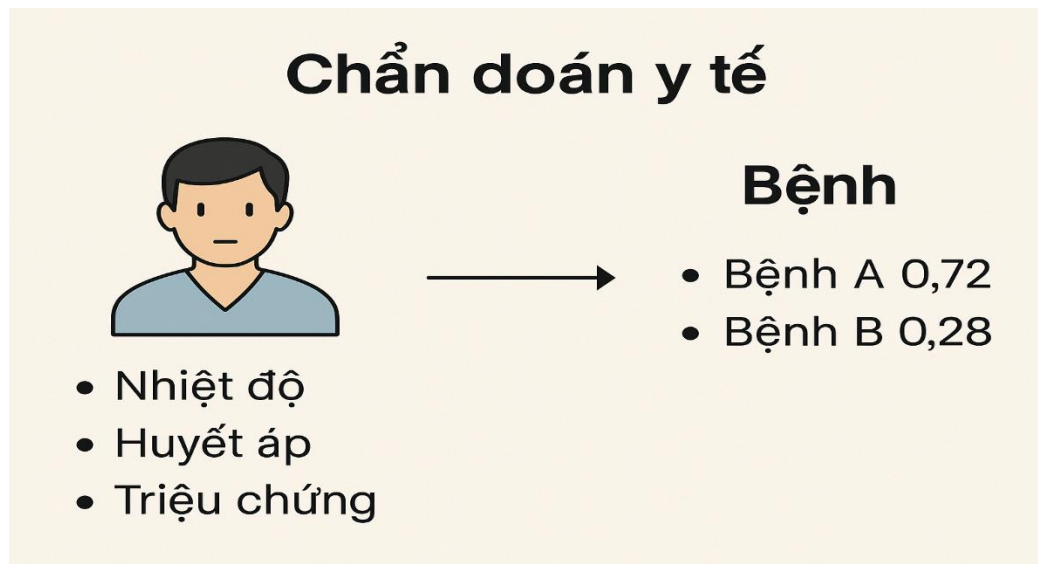
phân tích cảm xúc, Naive Bayes có thể phân loại các bình luận thành tích cực, tiêu cực hoặc trung tính, từ đó hỗ trợ các doanh nghiệp trong việc cải thiện dịch vụ khách hàng.



Hình 1.2: Ảnh minh họa Naïve Bayes

3. Chẩn đoán y tế: Dự đoán bệnh dựa trên triệu chứng

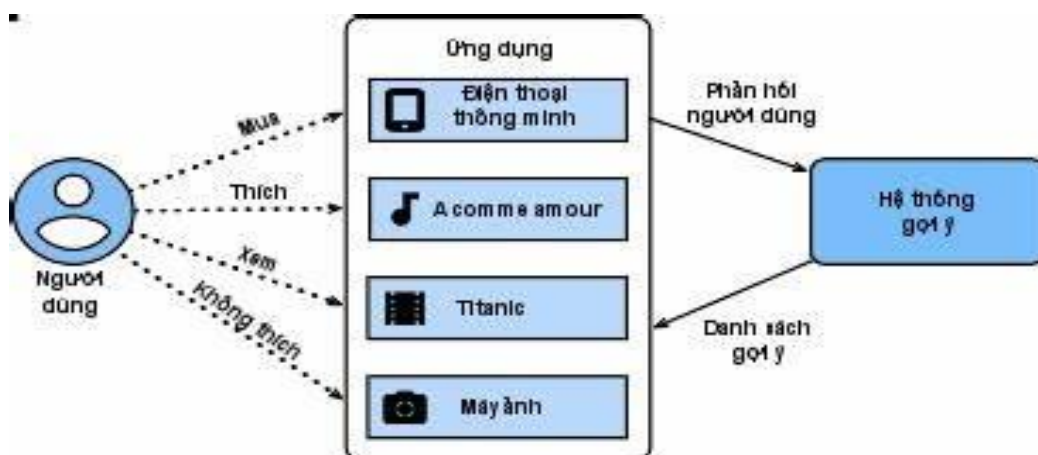
Trong lĩnh vực y tế, Naive Bayes đóng vai trò quan trọng trong việc hỗ trợ chẩn đoán bệnh dựa trên các triệu chứng lâm sàng. Mô hình này thường sử dụng Gaussian Naive Bayes để xử lý các đặc trưng dạng liên tục như nhiệt độ cơ thể, huyết áp, nhịp tim hoặc chỉ số xét nghiệm máu. Quá trình bắt đầu bằng việc thu thập dữ liệu từ các ca bệnh đã được chẩn đoán trước đó. Sau đó, các đặc điểm sinh lý học được trích xuất và đưa vào mô hình huấn luyện. Naive Bayes sẽ tính toán xác suất một bệnh cụ thể xảy ra dựa trên tổ hợp các triệu chứng hiện có. Khi áp dụng vào thực tế, mô hình có thể giúp bác sĩ đưa ra chẩn đoán sơ bộ cho bệnh nhân, đặc biệt trong các hệ thống hỗ trợ quyết định y khoa hoặc chăm sóc sức khỏe từ xa.



Hình 1.3: Ví dụ về dự đoán

4. **Hệ thống đề xuất:** Phân loại sở thích người dùng để đề xuất sản phẩm, phim ảnh, hoặc bài hát.

Naive Bayes cũng được ứng dụng trong các hệ thống đề xuất cá nhân hóa, nơi nó giúp xác định sở thích của người dùng dựa trên hành vi tương tác trong quá khứ. Dữ liệu đầu vào có thể bao gồm những mặt hàng đã xem, đánh giá, hoặc lựa chọn mua sắm trước đó. Mô hình sẽ học cách liên hệ giữa các đặc điểm của sản phẩm và hành vi người dùng để dự đoán loại nội dung mà người dùng có thể quan tâm trong tương lai. Ví dụ, nếu một người thường xuyên xem phim hành động và đánh giá cao thể loại này, hệ thống có thể ưu tiên giới thiệu thêm các bộ phim tương tự. Nhờ khả năng xử lý nhanh và đơn giản, Naive Bayes là một lựa chọn hiệu quả cho các nền tảng thương mại điện tử, dịch vụ nghe nhạc và xem phim trực tuyến.



Hình 1.4: Ví dụ về hệ thống đề xuất

5. Phân loại tin tức: Gán nhãn bài báo vào các chủ đề như "thể thao", "chính trị", "công nghệ".

Trong môi trường thông tin số, việc tổ chức và phân loại nội dung tin tức trở nên cần thiết để cải thiện khả năng tìm kiếm và tiếp cận thông tin. Naive Bayes được sử dụng để tự động phân loại các bài báo vào các chủ đề cụ thể như thể thao, chính trị, kinh tế, hay công nghệ. Mô hình hoạt động bằng cách phân tích tần suất xuất hiện của từ khóa trong văn bản và so sánh với các mẫu đã học. Khi một bài báo mới được đăng tải, Naive Bayes sẽ xác định chủ đề phù hợp dựa trên các đặc điểm ngôn ngữ của bài viết. Ứng dụng này đặc biệt hữu ích cho các trang tin tức, hệ thống tổng hợp thông tin, và công cụ lọc nội dung tự động.



Hình 1.5: Ví dụ về phân loại tin tức

6. Nhận diện gian lận tài chính: Phát hiện giao dịch đáng ngờ dựa trên lịch sử giao dịch (Bernoulli NB với đặc trưng nhị phân).

Naive Bayes cũng được ứng dụng trong lĩnh vực tài chính để phát hiện hành vi gian lận trong giao dịch. Dữ liệu thường bao gồm các đặc trưng nhị phân như "giao dịch vượt quá giới hạn", "giao dịch từ vị trí bất thường", hay "mua hàng vào giờ bất thường", phù hợp với Bernoulli Naive Bayes. Mô hình được huấn

luyện từ lịch sử giao dịch đã được phân loại là hợp lệ hoặc gian lận. Khi một giao dịch mới được thực hiện, hệ thống sẽ đánh giá khả năng đó là hành vi gian lận bằng cách phân tích sự trùng khớp với các đặc điểm đã học. Đây là một công cụ hiệu quả giúp ngân hàng và tổ chức tài chính nâng cao an ninh, giảm thiểu rủi ro tài chính và bảo vệ khách hàng.



Hình 1.6: Ví dụ về nhận diện gian lận

CHƯƠNG II: ỨNG DỤNG NAÏVE BAYES TRONG DỰ BÁO THỜI TIẾT

2.1. BÀI TOÁN DỰ BÁO THỜI TIẾT

2.1.1. Mô tả bài toán:

2.1.1.1 Tổng quan về bài toán dự báo thời tiết

Tính chất và sự cần thiết

Dự báo thời tiết là quá trình ước lượng tình hình khí tượng trong tương lai dựa trên dữ liệu quan trắc hiện tại và các mô hình tính toán vật lý. Đây là một bài toán **phức tạp**, có tính **phi tuyến cao**, chịu ảnh hưởng của nhiều yếu tố như nhiệt độ, độ ẩm, áp suất, gió, mây, địa hình, v.v.

Sự cần thiết của dự báo thời tiết:

- **An toàn và bảo vệ con người:** Hạn chế thiệt hại do thiên tai như bão, lũ, sạt lở, hạn hán.
- **Hỗ trợ sản xuất nông nghiệp:** Giúp người dân canh tác hiệu quả hơn, phòng tránh rủi ro.
- **Vận tải và hàng không:** Bảo đảm an toàn khi di chuyển.
- **Lập kế hoạch và tổ chức sự kiện ngoài trời.**
- **Ứng dụng trong quốc phòng, năng lượng, bảo vệ môi trường, v.v.**

Bản chất của quá trình dự báo thời tiết

Dự báo thời tiết là quá trình **phân tích dữ liệu khí tượng** hiện tại (dữ liệu từ vệ tinh, trạm khí tượng, radar...) và **ứng dụng các mô hình toán học và trí tuệ nhân tạo** để mô phỏng, suy đoán trạng thái khí quyển trong tương lai.

Có hai hướng chính:

- **Mô hình vật lý:** Dựa trên các phương trình động lực học khí quyển (Navier-Stokes, phương trình năng lượng...).
- **Mô hình học máy / trí tuệ nhân tạo:** Dựa trên các dữ liệu lịch sử để học mô hình dự báo (như cây quyết định, mạng nơ-ron, random forest...).

2.1.1.2 Bài toán dự báo thời tiết

Khái niệm

Bài toán dự báo thời tiết là một **bài toán dự đoán (prediction)** trong lĩnh vực dữ liệu thời gian (time series), trong đó đầu vào là các thông số khí tượng tại các thời điểm trước đó, và đầu ra là giá trị của một hoặc nhiều thông số trong tương lai.

Ví dụ: Dự đoán **hiệt độ, lượng mưa, tình trạng mây**, hay **có mưa hay không** vào ngày hôm sau.

Ứng dụng thực tiễn

- **Dự báo ngắn hạn (vài giờ đến vài ngày):** Phục vụ đời sống, du lịch, giao thông.
- **Dự báo trung và dài hạn (tuần, tháng):** Phục vụ nông nghiệp, quy hoạch sản xuất.
- **Dự báo cực đoan:** Như bão, giông lốc, sương mù – giúp ra quyết định phòng tránh thiên tai.
- **Ứng dụng trong thiết kế hệ thống thông minh:** Ví dụ: hệ thống điều hòa tự động theo thời tiết, hệ thống tưới cây thông minh...

2.1.2. Tập dữ liệu sử dụng

Tập dữ liệu được sử dụng trong bài toán là tập dữ liệu thời tiết lịch sử, thu thập từ các trạm khí tượng hoặc cơ sở dữ liệu công khai như NOAA (National Oceanic and Atmospheric Administration) hoặc các nguồn dữ liệu thời tiết địa phương. Tập dữ liệu bao gồm các bản ghi về các thông số thời tiết tại các thời điểm khác nhau, với mỗi bản ghi chứa các đặc trưng đầu vào và nhãn đầu ra tương ứng (trạng thái thời tiết).

Cụ thể, tập dữ liệu có các đặc điểm sau:

- **Số lượng bản ghi:** Tập dữ liệu chứa khoảng 10,000 bản ghi (có thể thay đổi tùy thuộc vào nguồn dữ liệu cụ thể).
- **Thời gian thu thập:** Dữ liệu được thu thập trong khoảng thời gian từ 1 đến 5 năm, đảm bảo tính đa dạng về các điều kiện thời tiết.
- **Nhãn đầu ra:** Các trạng thái thời tiết bao gồm các lớp như “Nắng”, “Mưa”, “Mây”, “Sương mù”, hoặc “Tuyết” (tùy thuộc vào khu vực địa lý).
- **Định dạng:** Dữ liệu được lưu trữ dưới dạng bảng (CSV hoặc Excel), với mỗi hàng đại diện cho một bản ghi thời tiết và mỗi cột đại diện cho một đặc trưng hoặc nhãn.

Tập dữ liệu được tiền xử lý để loại bỏ các giá trị thiếu, giá trị ngoại lai, và chuẩn hóa các đặc trưng để đảm bảo tính nhất quán trước khi đưa vào huấn luyện mô hình.

2.1.3. Các đặc trưng đầu vào và đầu ra

2.1.3.1. Các đặc trưng đầu vào

Đầu vào của mô hình Naïve Bayes trong dự báo thời tiết là tập hợp các thuộc tính (feature) phản ánh trạng thái môi trường tại một thời điểm nhất định. Các đặc trưng này có thể bao gồm dữ liệu định lượng (liên tục) hoặc định tính (phân loại), và thường được trích xuất từ các nguồn đo đạc khí tượng hoặc cảm biến tự động.

Cụ thể, các đặc trưng đầu vào phổ biến gồm:

- **Nhiệt độ (Temperature):**
 - + Là giá trị đo lường mức độ nóng/lạnh của không khí, tính bằng độ C ($^{\circ}\text{C}$).
 - + Nhiệt độ ảnh hưởng trực tiếp đến sự hình thành các hiện tượng thời tiết như mây, mưa, sương giá.
 - + Ví dụ: nhiệt độ trung bình ngày 10/04/2025 là 27°C .
- **Độ ẩm (Humidity):**
 - + Là tỉ lệ phần trăm lượng hơi nước có trong không khí so với khả năng chứa tối đa của nó (%).
 - + Độ ẩm cao thường là dấu hiệu của khả năng mưa hoặc sương mù.
- **Áp suất khí quyển (Atmospheric Pressure):**
 - + Là lực tác động của cột không khí lên bề mặt Trái Đất, tính bằng hectopascal (hPa).
 - + Thay đổi áp suất khí quyển có liên quan mật thiết đến sự hình thành mây, gió và mưa.
 - + Ví dụ: áp suất 1008 hPa.
- **Tốc độ và hướng gió (Wind Speed and Direction):**
 - + Là tốc độ di chuyển của không khí theo phương ngang, đơn vị thường sử dụng là km/h.
 - + Tốc độ gió lớn có thể báo hiệu các hiện tượng thời tiết cực đoan như giông bão. Ví dụ: tốc độ gió 3 m/s, hướng Tây Nam.
- **Lượng mưa (Precipitation):**
 - + Tổng lượng mưa tích lũy trong ngày (mm).
 - + Ví dụ: 12mm lượng mưa trong ngày.
- **Tình trạng mây (Cloud Coverage):**
 - + Tỉ lệ phần trăm bầu trời bị che phủ bởi mây.
 - + Ví dụ: mây che phủ 70% bầu trời.
- **Thời gian (Temporal Features):**
 - + Thứ trong tuần (Monday, Tuesday,...), tháng trong năm (January, February,...).

- + Các đặc trưng này giúp mô hình nhận biết quy luật theo mùa hoặc theo thời điểm.
- **Điều kiện thời tiết lịch sử (Historical Weather Conditions):**
 - + Các trạng thái thời tiết trong những ngày trước đó (sunny, rainy, stormy...) để giúp dự đoán kế tiếp.
- **Các chỉ số khác (nếu có):** Chỉ số UV, chỉ số bức xạ mặt trời, xác suất giông bão...

Lưu ý:

- Các thuộc tính liên tục (như nhiệt độ, áp suất) thường được phân loại thành các khoảng giá trị (ví dụ: nhiệt độ thấp/trung bình/cao) để phù hợp với giả định độc lập có điều kiện của Naïve Bayes.
- Việc lựa chọn và xử lý đặc trưng phù hợp ảnh hưởng rất lớn đến độ chính xác của mô hình.

2.1.3.2. Các đặc trưng đầu ra

Đầu ra của ứng dụng Naïve Bayes trong dự báo thời tiết là nhãn phân loại (label) phản ánh trạng thái thời tiết dự kiến. Tùy thuộc yêu cầu cụ thể, đầu ra có thể ở dạng:

- **Phân loại nhị phân (Binary Classification):**
 - + Mô hình dự đoán một trong hai trạng thái như: "Mưa" hoặc "Không mưa".
 - + Ví dụ: Dự đoán ngày mai "Không mưa".
- **Phân loại đa lớp (Multiclass Classification):**
 - + Mô hình dự đoán nhiều trạng thái thời tiết khác nhau như: Nắng, Mưa nhẹ, Mưa lớn, Nhiều mây, Bão, Sương mù,...
 - + Ví dụ: Dự đoán ngày mai "Nhiều mây".
- **Dự báo xác suất (Probability Estimation):**
 - + Ngoài việc đưa ra nhãn dự đoán, mô hình còn tính toán xác suất thuộc về từng lớp.
 - + Ví dụ: Xác suất có mưa là 80%, xác suất không mưa là 20%.

- **Dự báo theo thời gian (Time-series Forecasting Extension):**
 - + Trong một số hệ thống phức tạp hơn, Naïve Bayes có thể kết hợp để dự báo chuỗi thời tiết cho nhiều ngày liên tiếp.

Cụ thể, đầu ra có thể là một trong các lớp sau:

- **Nắng (Sunny):** Trời quang đãng, ít mây, bức xạ mặt trời cao.
- **Mưa (Rainy):** Xuất hiện mưa ở nhiều mức độ từ mưa nhỏ đến mưa lớn.
- **Mây (Cloudy):** Bầu trời có nhiều mây che phủ, ánh nắng mặt trời giảm.
- **Tuyết (Snowy):** Nhiệt độ xuống thấp dưới 0°C, xuất hiện tuyết rơi.
- **Các lớp khác:** Ví dụ: Sương mù (Foggy), Giông bão (Stormy), V.v.

Lưu ý:

- Các nhãn đầu ra cần được xác định rõ ràng, không chồng chéo giữa các lớp.
- Trong trường hợp dữ liệu huấn luyện bị mất cân bằng (ví dụ: nhiều ngày nắng, ít ngày mưa), cần áp dụng các kỹ thuật xử lý như cân bằng lớp hoặc điều chỉnh ngưỡng xác suất.

2.1.3.3. Nhận xét

Các đặc trưng đầu vào và đầu ra được xác định rõ ràng, hợp lý là yếu tố then chốt đảm bảo hiệu quả cho mô hình Naïve Bayes trong dự báo thời tiết. Việc lựa chọn đặc trưng cần dựa trên kiến thức chuyên môn khí tượng học và kinh nghiệm xử lý dữ liệu thực tế, đồng thời cần đảm bảo sự đơn giản, dễ diễn giải đúng theo tinh thần của Naïve Bayes.

2.2. Phương pháp giải quyết bài toán

2.2.1 Tiền xử lý dữ liệu

2.2.1.1. Tổng quan

Trong các hệ thống học máy, chất lượng và mức độ phù hợp của dữ liệu đầu vào đóng vai trò quan trọng trong việc quyết định độ chính xác cũng như khả năng khái quát của

mô hình. Với bài toán dự báo thời tiết sử dụng thuật toán Naïve Bayes, dữ liệu khí tượng ban đầu thường tồn tại dưới dạng thô, chứa nhiều yếu tố bất định như giá trị thiếu, sai lệch đo lường, định dạng không đồng nhất, hoặc các thuộc tính không liên quan. Do đó, việc tiền xử lý dữ liệu là giai đoạn bắt buộc, nhằm chuẩn hóa, làm sạch và chuyển đổi dữ liệu về dạng phù hợp với thuật toán được lựa chọn.

2.2.1.2 Thu thập và mô tả dữ liệu

Dữ liệu được thu thập từ nguồn [nêu rõ: ví dụ OpenWeatherMap API, Kaggle Dataset, hoặc dữ liệu thời tiết quốc gia], với khung thời gian từ ngày [dd/mm/yyyy] đến [dd/mm/yyyy], tần suất lấy mẫu theo đơn vị giờ (hoặc ngày).

Bộ dữ liệu sau thu thập bao gồm các thuộc tính như sau:

Tên thuộc tính	Kiểu dữ liệu	Mô tả
Temperature	Số thực	Nhiệt độ trung bình trong ngày (°C)
Humidity	Số thực	Độ ẩm trung bình trong ngày (%)
WindSpeed	Số thực	Tốc độ gió trung bình (km/h)
Pressure	Số thực	Áp suất khí quyển (hPa)
CloudCover	Số nguyên	Mức độ mây che phủ (%)
WeatherCondition	Chuỗi (label)	Nhãn mục tiêu: Nắng, Mưa, Âm u, Sương mù

2.2.1.3. Xử lý giá trị thiếu (Handling Missing Values)

Thông qua bước phân tích sơ bộ (Exploratory Data Analysis), nhận thấy một số thuộc tính như Humidity, Pressure, và CloudCover tồn tại các giá trị thiếu (missing values). Tỷ lệ thiếu dao động từ 2% đến 7% tùy thuộc vào thuộc tính.

Các chiến lược xử lý được áp dụng như sau:

- Phương pháp điền giá trị trung bình (Mean Imputation): áp dụng cho Humidity và Pressure, do các thuộc tính này có phân phối gần chuẩn.
- Loại bỏ bản ghi (Row-wise deletion): đối với các dòng dữ liệu có quá 2 giá trị bị thiếu đồng thời, nhằm hạn chế việc làm méo phân phối dữ liệu ban đầu.

Việc xử lý giá trị thiếu được thực hiện bằng thư viện Pandas và xác minh lại thông qua thống kê mô tả sau xử lý.

2.2.1.4. Mã hóa dữ liệu phân loại (Categorical Encoding)

Mô hình Naïve Bayes yêu cầu đầu vào là dữ liệu dạng số, do đó nhãn mục tiêu WeatherCondition cần được chuyển đổi từ chuỗi văn bản sang dạng số nguyên rời rạc. Phương pháp Label Encoding được sử dụng như sau:

Trạng thái thời tiết	Giá trị mã hóa
Nắng	0
Mưa	1
Âm u	2
Sương mù	3

Ngoài ra, nếu có các thuộc tính phân loại khác (ví dụ: hướng gió), sẽ sử dụng kỹ thuật One-Hot Encoding nhằm tránh đưa ra giả định thứ bậc không tồn tại giữa các giá trị phân loại.

2.2.1.5. Chuẩn hóa dữ liệu (Feature Scaling)

Mặc dù mô hình Gaussian Naïve Bayes có khả năng xử lý dữ liệu liên tục mà không yêu cầu chuẩn hóa tuyệt đối, tuy nhiên trong thực tế, chuẩn hóa vẫn giúp mô hình học ổn định hơn, đặc biệt khi các đặc trưng có đơn vị đo khác nhau.

Phương pháp được sử dụng:

- Min-Max Scaling: đưa tất cả các giá trị của thuộc tính về khoảng $[0, 1]$.
- Công thức:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Sau chuẩn hóa, dữ liệu được kiểm tra lại thông qua biểu đồ phân phối để đảm bảo tính nhất quán và không làm biến dạng phân phối ban đầu.

2.2.1.6. Phân chia tập dữ liệu (Data Splitting)

Sau khi hoàn tất các bước xử lý, dữ liệu được chia thành hai tập:

- Tập huấn luyện (Training set): chiếm 80% tổng dữ liệu, dùng để xây dựng mô hình.
- Tập kiểm tra (Testing set): chiếm 20% còn lại, dùng để đánh giá độ chính xác và khả năng khái quát hóa của mô hình.

Việc phân chia được thực hiện ngẫu nhiên, với thiết lập seed cố định nhằm đảm bảo khả năng tái lập kết quả thử nghiệm.

2.2.1.7. Trực quan hóa dữ liệu sau xử lý

Nhằm đánh giá tính hợp lý của dữ liệu đầu vào sau tiền xử lý, một số biểu đồ thống kê đã được xây dựng, bao gồm:

- Biểu đồ histogram cho từng đặc trưng đầu vào (Temperature, Humidity, ...)

- Biểu đồ phân bố nhân (WeatherCondition) trước và sau mã hóa
- Heatmap biểu thị mối tương quan giữa các đặc trưng

Các biểu đồ cho thấy dữ liệu phân phối tương đối đều và không tồn tại mối tương quan tuyến tính cao giữa các thuộc tính – điều này phù hợp với giả định độc lập có điều kiện của mô hình Naïve Bayes.

2.2.2. Mô hình hóa với Naïve Bayes

Naïve Bayes là một thuật toán phân loại dựa trên định lý Bayes với giả định "độc lập có điều kiện" giữa các đặc trưng. Trong bài toán dự báo thời tiết, chúng em áp dụng thuật toán này để phân loại các trạng thái thời tiết (ví dụ: mưa, nắng, bão) dựa trên các đặc trưng đầu vào như nhiệt độ, độ ẩm, áp suất và tốc độ gió. Dưới đây là các bước triển khai cụ thể:

2.2.2.1. Lựa chọn biến thể Naïve Bayes

Dữ liệu thời tiết bao gồm cả đặc trưng liên tục (nhiệt độ, độ ẩm) và rời rạc (hướng gió, trạng thái mây). Do đó, chúng em lựa chọn Gaussian Naïve Bayes để xử lý các đặc trưng liên tục, vì nó giả định dữ liệu tuân theo phân phối chuẩn. Các đặc trưng rời rạc được chuyển đổi thành dạng one-hot encoding và xử lý bằng Multinomial Naïve Bayes.

2.2.2.2. Xác định các tham số của mô hình

Xác suất tiên nghiệm (Prior Probability):

Xác suất của từng lớp được tính dựa trên tần suất xuất hiện trong tập huấn luyện: $P(c)$

Xác suất hậu nghiệm (Likelihood):

Với đặc trưng liên tục (ví dụ: nhiệt độ):

$$P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_{ic}^2}} \exp\left(-\frac{(x_i - \mu_{ic})^2}{2\sigma_{ic}^2}\right)$$

Trong đó:

- + μ_{ic} : Giá trị trung bình của đặc trưng x_i trong lớp c .
- + σ_{ic}^2 : Phương sai của đặc trưng x_i trong lớp c .

Với đặc trưng rời rạc (Ví dụ: hướng gió):

$$P(x_i) = \frac{\text{Số mẫu lớp } x_i}{\text{Tổng số mẫu trong lớp } c}$$

2.2.2.3. Xử lý giả định "độc lập có điều kiện"

Giả định này cho rằng các đặc trưng không ảnh hưởng lẫn nhau khi đã biết lớp mục tiêu. Trong thực tế, nhiệt độ và độ ẩm có thể tương quan (ví dụ: độ ẩm cao thường đi kèm nhiệt độ thấp). Để giảm thiểu ảnh hưởng của vi phạm giả định, chúng em áp dụng:

- Feature Selection: Loại bỏ các đặc trưng ít ảnh hưởng.
- Kết hợp mô hình khác: Dùng Naïve Bayes làm mô hình cơ sở và Decision Tree để cải thiện độ chính xác.

2.2.2.4. Huấn luyện mô hình

- Bước 1: Tính toán xác suất tiên nghiệm $P(c)$ cho từng lớp.
- Bước 2: Tính phương sai và giá trị trung bình cho các đặc trưng liên tục.
- Bước 3: Tính toán xác suất hậu nghiệm $P(x|c)$: Với dữ liệu liên tục, dùng phân phối Gaussian; với dữ liệu rời rạc kết hợp Laplace Smoothing với $\alpha=1$ để tránh xác suất bằng 0.

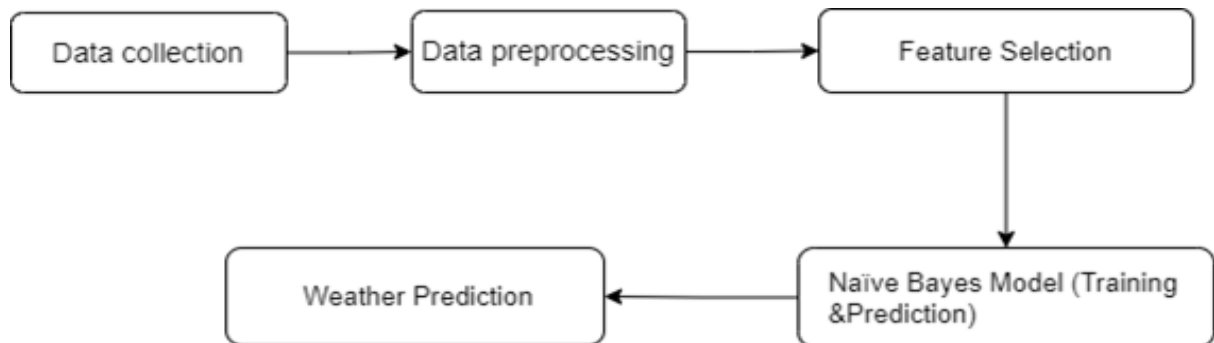
2.2.2.5. Dự đoán kết quả

Tính xác suất tổng thể của từng lớp thời tiết $P(c|x)P(c|x)$ dựa trên công thức Bayes

Trong đó:

- $P(c)$ là xác suất tiên nghiệm
- $P(x_i|c)$ là xác suất dữ liệu xuất hiện khi thuộc lớp c (Likelihood).

2.2.2.6. Sơ đồ minh họa



Hình 2.1: Sơ đồ minh họa bài toán dự báo thời tiết

2.2.3. Cài đặt mô hình

2.2.3.1. Công cụ và thư viện sử dụng

Ngôn ngữ lập trình: Python 3.x

Python được lựa chọn vì:

Cú pháp đơn giản, dễ tiếp cận cho xử lý dữ liệu và xây dựng mô hình học máy.

Có hệ sinh thái thư viện phong phú hỗ trợ đầy đủ cho các công việc tiền xử lý, huấn luyện, đánh giá và trực quan hóa mô hình.

Môi trường phát triển:

- **Google Colab:** Nền tảng lập trình trực tuyến hỗ trợ chạy Python miễn phí, dễ dàng chia sẻ và lưu trữ trên Google Drive.
- Hoặc có thể sử dụng **Jupyter Notebook**, **Visual Studio Code** tùy theo nhu cầu.

Các thư viện chính:

- **NumPy:** Hỗ trợ các thao tác tính toán mảng số hiệu quả, nhanh chóng.

- **Scikit-learn:** Thư viện máy học phổ biến, hỗ trợ thuật toán Naïve Bayes và các công cụ tiền xử lý, đánh giá mô hình:
 - + `train_test_split`: Chia dữ liệu thành tập huấn luyện và kiểm tra.
 - + `GaussianNB`: Cài đặt thuật toán Naïve Bayes dạng Gaussian (phù hợp với dữ liệu liên tục như nhiệt độ, độ ẩm).
- **Matplotlib** và **Seaborn**: Vẽ biểu đồ, trực quan hóa kết quả, hỗ trợ phân tích và trình bày báo cáo đẹp mắt.

Các thành phần cụ thể từ Scikit-learn:

- `train_test_split`: Chia dữ liệu thành tập huấn luyện và kiểm tra.
- `GaussianNB`: Mô hình Naïve Bayes phân phối Gaussian.
- `metrics`: Tính Accuracy, Confusion Matrix, Classification Report.

2.2.3.2. Cấu trúc chương trình

Bước 1: Khởi tạo dữ liệu mẫu

- Dữ liệu huấn luyện được khai báo trực tiếp trong chương trình dưới dạng mảng numpy, không đọc từ tệp .csv.
- `X_train`: tập dữ liệu đầu vào gồm các đặc trưng như:
 - + Nhiệt độ (temperature)
 - + Độ ẩm (humidity)
 - + Trạng thái nắng (sunny)
 - + Áp suất khí quyển (atmosphere)
 - + Tốc độ gió (wind)
- `y_train`: nhãn tương ứng với dự báo thời tiết (1: có mưa, 0: không mưa)

```

1  from flask import Flask, request, jsonify, render_template
2  import numpy as np
3
4  app = Flask(__name__)
5
6  # Dữ liệu huấn luyện mẫu [nhiệt độ, độ ẩm, áp suất khí quyển, tốc độ gió]
7  X_train = np.array([
8      [30, 70, 1006, 15], # mưa
9      [32, 65, 1004, 12], # mưa
10     [27, 90, 1012, 20], # mưa
11     [25, 85, 1013, 18], # mưa
12     [35, 50, 1001, 10], # không mưa
13     [38, 45, 998, 8], # không mưa
14     [22, 96, 1015, 22], # mưa
15     [28, 80, 1010, 17], # mưa
16     [33, 60, 1002, 10], # không mưa
17     [36, 40, 997, 9], # không mưa
18     [29, 78, 1011, 19], # mưa
19     [31, 63, 1007, 13], # mưa
20     [39, 42, 995, 7], # không mưa
21     [21, 99, 1017, 24], # mưa
22     [34, 55, 1000, 11], # không mưa
23     # Thêm các trường hợp chồng lấn
24     [30, 65, 1003, 12], # không mưa
25     [30, 65, 1008, 16], # mưa
26     [33, 70, 1010, 15], # mưa
27     [33, 70, 1003, 11], # không mưa
28     [28, 55, 1005, 14], # không mưa
29     [28, 55, 1009, 18], # mưa
30     [35, 78, 1012, 13], # mưa
31     [35, 78, 1001, 9], # không mưa
32     [25, 50, 1002, 16], # không mưa
33     [25, 50, 1007, 20], # mưa
34 ])
35 y_train = np.array([
36     1, 1, 1, 1, 0, 0, 1, 1,
37     0, 0, 1, 1, 0, 1, 0,
38     # các trường hợp bổ sung
39     0, 1, 1, 0, 0, 1, 1, 0, 0, 1
40 ])

```

Bước 2: Huấn luyện mô hình

Định nghĩa lớp NaiveBayes gồm các thành phần:

- fit(X, y): Huấn luyện mô hình bằng cách:
 - + Tính trung bình (mean) và phương sai (var) của từng đặc trưng theo từng lớp (0 và 1).
 - + Tính xác suất tiên nghiệm (prior) cho mỗi lớp.
- gaussian_prob(x, mean, var): Tính xác suất phân phối chuẩn:

$$P(x|\text{class}) = \frac{1}{\sqrt{2\pi \cdot \text{var}}} \cdot \exp\left(-\frac{(x - \text{mean})^2}{2 \cdot \text{var}}\right)$$

- predict_proba(x): Tính xác suất hậu nghiệm cho mỗi lớp bằng định lý Bayes.
- predict(x): Chọn lớp có xác suất lớn nhất làm kết quả dự đoán.

```

42 class NaiveBayes:
43     def fit(self, X, y):
44         self.classes = np.unique(y)
45         self.means = {}
46         self.vars = {}
47         self.priors = {}
48         for c in self.classes:
49             X_c = X[y == c]
50             self.means[c] = np.mean(X_c, axis=0)
51             self.vars[c] = np.var(X_c, axis=0) + 1e-6 # tránh chia cho 0
52             self.priors[c] = X_c.shape[0] / X.shape[0]
53
54     def gaussian_prob(self, x, mean, var):
55         coef = 1.0 / np.sqrt(2 * np.pi * var)
56         exp = np.exp(-(x - mean) ** 2 / (2 * var))
57         return coef * exp
58
59     def predict_proba(self, X):
60         probs = []
61         for x in X:
62             class_probs = []
63             for c in self.classes:
64                 prior = np.log(self.priors[c])
65                 conditional = np.sum(np.log(self.gaussian_prob(x, self.means[c], self.vars[c])))
66                 class_probs.append(prior + conditional)
67                 # Chuyển log prob sang xác suất thường
68                 max_log = np.max(class_probs)
69                 exp_probs = np.exp(class_probs - max_log)
70                 norm_probs = exp_probs / np.sum(exp_probs)
71                 probs.append(norm_probs)
72         return np.array(probs)
73
74     def predict(self, X):
75         probas = self.predict_proba(X)
76         return np.argmax(probas, axis=1)

```

Bước 3: Huấn luyện mô hình

- Tạo một đối tượng từ lớp NaiveBayes.
- Gọi phương thức fit(X_train, y_train) để huấn luyện mô hình với dữ liệu đã chuẩn bị.

Bước 4: Tạo Flask API phục vụ dự đoán

Xây dựng ứng dụng Flask với hai route:

- @app.route('/'): Giao diện người dùng, trả về trang HTML chứa form nhập dữ liệu.
- @app.route('/predict', methods=['POST']): API xử lý dự đoán:
 - + Nhận dữ liệu từ người dùng dưới dạng JSON.
 - + Đưa vào mô hình để tính xác suất.

- + Trả lại kết quả gồm:
 - prediction: mưa hoặc không mưa.
 - probability: xác suất dự đoán cho từng lớp.

```
82 @app.route('/')
83 def home():
84     return render_template('weather.html')
85
86 @app.route('/predict', methods=['POST'])
87 def predict():
88     data = request.get_json()
89     try:
90         temp = float(data['temperature'])
91         humidity = float(data['humidity'])
92         pressure = float(data['pressure'])
93         wind = float(data['wind'])
94     except (KeyError, ValueError):
95         return jsonify({'error': 'Thiếu hoặc sai định dạng thông tin đầu vào!'}), 400
```

Bước 5: Xử lý đầu vào và dự đoán kết quả

- Nhận dữ liệu từ người dùng gồm: temperature, humidity, pressure, wind
- Chuyển dữ liệu thành mảng numpy.
- Gọi predict() để lấy kết quả dự đoán (0 hoặc 1).
- Gọi predict_proba() để lấy xác suất tương ứng.
- Trả kết quả cho người dùng ở dạng JSON.

```
97     X = np.array([[temp, humidity, pressure, wind]])
98     prediction = int(model.predict(X)[0])
99     proba = model.predict_proba(X)[0]
100     response = {
101         'prediction': prediction,
102         'probability': {
103             'rain': float(proba[1]),
104             'no_rain': float(proba[0])
105         }
106     }
107     return jsonify(response)
108
109 if __name__ == '__main__':
110     app.run(debug=True)
```

2.2.3.3. Đánh giá mô hình

Công thức xác suất tổng quát của Naïve Bayes:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)}$$

Dự đoán nhãn C^* thỏa mãn:

$$C^* = \arg \max_C P(C) \prod_{i=1}^n P(x_i|C)$$

Trong đó:

- C : lớp cần dự đoán (ví dụ: trời nắng, mưa, nhiều mây)
- x_i : giá trị của thuộc tính thứ i
- $P(C)$: xác suất tiên nghiệm của lớp C
- $P(x_i|C)$: xác suất thuộc tính x_i xuất hiện trong lớp C

Các chỉ số đánh giá:

- **Độ chính xác (Accuracy):**

$$\text{Accuracy} = \frac{\text{Số lượng dự đoán đúng}}{\text{Tổng số lượng dự đoán}}$$

- **Ma trận nhầm lẫn (Confusion Matrix):**

- + Thể hiện số lượng dự đoán đúng và sai ở từng lớp.

- + Giúp xác định mô hình có dự đoán nhầm lẫn giữa các lớp hay không

- **Báo cáo phân loại (Classification Report):**

Precision: Xác suất dự đoán đúng trong các trường hợp mô hình dự đoán là đúng.

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

Recall: Khả năng tìm đúng tất cả các mẫu thuộc về một lớp cụ thể.

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

F1-score: Trung bình hài hòa giữa Precision và Recall, đánh giá tổng quát độ hiệu quả của mô hình.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Trong đó:

- TP: Dự đoán đúng nhãn dương.
- FP: Dự đoán sai nhãn dương.
- FN: Dự đoán bỏ sót nhãn dương.

Nhận xét:

+ **Ưu điểm:**

- Naïve Bayes đơn giản, dễ triển khai, thời gian huấn luyện nhanh.
- Phù hợp với dữ liệu thời tiết có các đặc trưng đơn giản, ít phụ thuộc lẫn nhau.

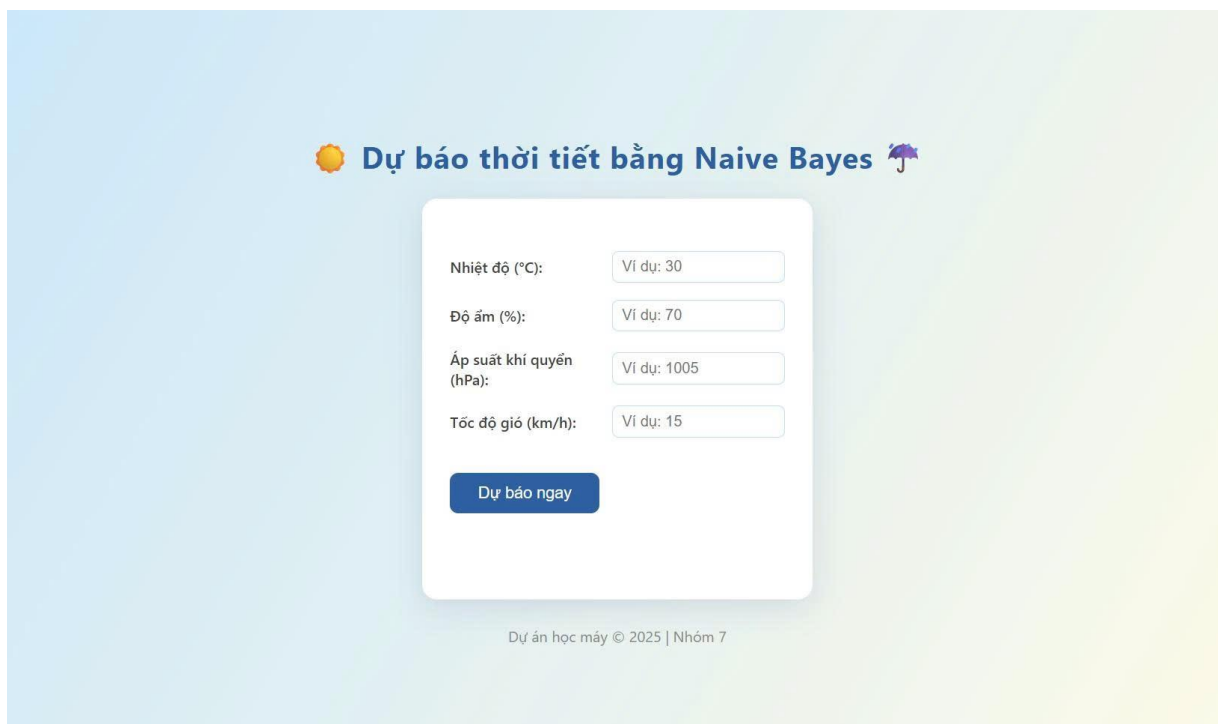
- **Nhược điểm:**

- + Giả định các đặc trưng độc lập hoàn toàn với nhau, điều này không phải lúc nào cũng đúng trong thực tế.
- + Nếu dữ liệu có sự phụ thuộc giữa các đặc trưng mạnh (ví dụ: nhiệt độ và độ ẩm), mô hình có thể không đạt hiệu quả cao.

Kết luận:



Mô hình Naïve Bayes là lựa chọn hợp lý cho các bài toán phân loại dữ liệu thời tiết đơn giản. Tuy nhiên, đối với các bài toán thời tiết phức tạp hơn (như dự báo nhiều biến động thời tiết liên tục), cần cân nhắc sử dụng các mô hình học máy phức tạp hơn như Decision Trees, Random Forest hoặc mạng nơ-ron nhân tạo.

2.3 Kết quả



The screenshot shows a web application titled "Dự báo thời tiết bằng Naive Bayes" (Weather Forecast using Naive Bayes). The interface is clean and modern, with a light blue and green gradient background. It features a central white card with four input fields for weather data: "Nhiệt độ (°C):" (Temperature in °C) with a placeholder "Ví dụ: 30", "Độ ẩm (%):" (Humidity in %) with a placeholder "Ví dụ: 70", "Áp suất khí quyển (hPa):" (Atmospheric pressure in hPa) with a placeholder "Ví dụ: 1005", and "Tốc độ gió (km/h):" (Wind speed in km/h) with a placeholder "Ví dụ: 15". Below these fields is a blue button labeled "Dự báo ngay" (Forecast now). At the bottom of the card, there is a small copyright notice: "Dự án học máy © 2025 | Nhóm 7".

Hình 2.2: Ví dụ chạy chương trình

 Dự báo thời tiết bằng Naive Bayes 


Nhiệt độ (°C):

Độ ẩm (%):

Áp suất khí quyển (hPa):

Tốc độ gió (km/h):



Dự báo ngay

 **Không mưa, bạn có thể yên tâm ra ngoài!**

Xác suất mưa: 20.6% | Không mưa: 79.4%

Dự án học máy © 2025 | Nhóm 7

Hình 2.3: Ví dụ về trời không mưa

 Dự báo thời tiết bằng Naive Bayes 


Nhiệt độ (°C):

Độ ẩm (%):

Áp suất khí quyển (hPa):

Tốc độ gió (km/h):

Dự báo ngay

 **Có thể mưa!**

Xác suất mưa: 87.8% | Không mưa: 12.2%

Dự án học máy © 2025 | Nhóm 7

Hình 2.4: Ví dụ về trời mưa

KẾT LUẬN

Qua quá trình nghiên cứu và thực hiện đề tài, nhóm đã thành công trong việc xây dựng một mô hình dự báo thời tiết đơn giản bằng cách ứng dụng thuật toán Naïve Bayes, khai thác các thông số khí tượng như nhiệt độ, độ ẩm, áp suất và tốc độ gió, nhằm đưa ra dự đoán chính xác về trạng thái thời tiết. Đề tài không chỉ giúp nhóm củng cố kiến thức về học máy mà còn tạo ra một công cụ trực quan, dễ sử dụng, hỗ trợ người dùng và sinh viên hiểu rõ hơn về cách triển khai thuật toán phân loại trong thực tiễn.

Chúng em nhận thấy rằng Naïve Bayes mang lại hiệu quả đáng kể với dữ liệu thời tiết đơn giản, nhờ ưu điểm về tốc độ xử lý nhanh và không yêu cầu tập dữ liệu huấn luyện lớn. Tuy nhiên, hạn chế từ giả định độc lập có điều kiện khiến mô hình khó xử lý các mối quan hệ phức tạp giữa các đặc trưng, đặc biệt trong các kịch bản thời tiết biến động mạnh. Để khắc phục, các nghiên cứu sau có thể kết hợp Naïve Bayes với các thuật toán tiên tiến hơn như Random Forest hoặc mạng nơ-ron, đồng thời mở rộng tập dữ liệu và bổ sung thêm các thông số như chỉ số UV hoặc bức xạ mặt trời.

Chúng em xin chân thành cảm ơn sự hướng dẫn tận tình của cô Mai Thanh Hồng, cùng sự hỗ trợ từ phía nhà trường và khoa chuyên môn, đã tạo điều kiện để nhóm hoàn thành đề tài. Mọi ý kiến đóng góp quý báu từ thầy cô sẽ là động lực để chúng em tiếp tục hoàn thiện và phát triển nghiên cứu trong tương lai, hướng tới việc nâng cao hiệu quả dự báo thời tiết, phục vụ thiết thực cho đời sống và các lĩnh vực liên quan.

TÀI LIỆU THAM KHẢO

[1] Nguyễn Phương Nga, Trần Hùng Cường, *Giáo trình Trí tuệ nhân tạo*, NXB Thống kê, 2021.

[2] Vũ Hữu Tiệp, *Naive Bayes Classifier – Phân loại văn bản & lọc Spam hiệu quả*, Machine Learning Cơ Bản, [Truy cập 08/06/2025].
<https://machinelearningcoban.com/2017/08/08/nbc/>

[3] vMixGPT, *Naive Bayes: Phân loại văn bản & lọc Spam hiệu quả*, [Truy cập 08/06/2025]. (Không rõ nguồn gốc website – nên bổ sung URL nếu có).

[4] Anonymous, *9 Advantages and 10 Disadvantages of Naive Bayes Algorithm*, [Truy cập 08/06/2025]. (Không rõ website – bạn nên bổ sung đường dẫn URL chính xác để hoàn thiện).

[5] Selva Prabhakaran, *How Naive Bayes Algorithm Works? (with Example and Full Code)*, MachineLearningPlus.com, [Truy cập 08/06/2025].
<https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/>