TRƯỜNG ĐẠI HỌC LẠC HỒNG

KHOA SAU ĐẠI HỌC

****

BÁO CÁO MÔN HỌC
THỐNG KÊ VÀ ỨNG DỤNG

ĐỀ TÀI:

Dự đoán hành vi tham gia chơi golf của khách hàng dựa trên dữ liệu thời tiết sử dụng Bernoulli Naive Bayes (BNB)

Giảng viên hướng dẫn:TS. Đỗ Sĩ Trường

Sinh viên thực hiện:

Đỗ Nguyễn Anh Tuấn - 923000224 - Trưởng nhóm

Nguyễn Thị Thuỳ - 923000211 - Phó nhóm

Nguyễn Ngọc Thái - 923000098

Cao Thế Phương - 923000031

Biên hòa, 06/2024

LỜI CẢM ƠN

Sau quá trình học tập và rèn luyện môn Thống kê và Ứng dụng tại trường Đại học Lạc Hồng dưới sự hướng dẫn của thầy Đỗ Sĩ Trường, chúng em đã được trang bị nhiều kiến thức bổ ích và các kỹ năng cần có để hoàn thành đề tài môn học của mình.

Xin chân thành gửi lời cảm ơn tới TS. Đỗ Sĩ Trường đã hướng dẫn, truyền đạt kiến thức và kinh nghiệm cho chúng em trong suốt thời gian học tập môn Thống kê và ứng dụng.

Trong quá trình làm đề tài kết thúc môn học không tránh khỏi được những sai sót, chúng em mong nhận được sự góp ý của thầy và các bạn để được hoàn thiện tốt hơn.

TP. Biên Hòa, tháng 6 năm 2024

MỤC LỤC

BẢNG THUẬT NGỮ ANH - VIỆT

Tiếng anh	Viết tắt	Tiếng việt
Bernoulli Naive Bayes	BNB	Thuật toán phân loại Bernoulli Naive Bayes
Maximum A Posteriori	MAP	Phương pháp MAP
Multinomial Naive Bayes	MultinomialNB	Naive Bayes đa thức
Gaussian Naive Bayes	GaussianNB	Thuật toán Gaussian Naive Bayes
Customer relationship management	CRM	Quản lý quan hệ khách hàng
Outlook		Thời tiết
Temperature	Temp	Nhiệt độ
Humidity	Hum	Độ ẩm
Windy		Gió
Prior Probability		Xác suất tiên nghiệm
Conditional Probability		Xác suất điều kiện
Precision	P	Độ chính xác
Recall	R	Độ nhạy
F1 Score	F1	Trung bình điều hòa giữa độ chuẩn xác và độ phủ

BẢNG CÁC KÝ HIỆU

Ký hiệu từ viết tắt	Diễn giải
P(x|y)	Xác suất của x khi biết y
P(y|x)	Xác suất của y khi x đã xảy ra
P(x)	Xác suất của sự kiện x
P(y)	Xác suất của sự kiên y
Π	Tích
θyi	Xác suất P(xi | y) của đặc trưng i xuất hiện trong một mẫu thuộc lớp y
θy	Vector theta cho lớp y chứa các xác suất θyi cho tất cả các đặc trưng
P(x1, ..., xn)	Xác suất của tất cả các biến x1 đến xn
n	Số lượng đặc trưng (features)
θ	Tham số theta
P(A∩B)	Xác suất của cả hai sự kiện A và B cùng xảy ra.

DANH MỤC CÁC HÌNH VÀ BẢNG

CHƯƠNG 1. MỞ ĐẦU

1. Bối cảnh nghiên cứu

Trong môi trường kinh doanh ngày nay, sự hiểu biết sâu sắc về khách hàng không chỉ là yếu tố then chốt mà còn là chìa khóa quyết định cho sự thành công của mọi doanh nghiệp. Dữ liệu khách hàng trở thành tài nguyên quý báu, là nguồn thông tin không thể phủ nhận về hành vi tiêu dùng, sở thích, và nhu cầu của họ. Trong bối cảnh thị trường cạnh tranh khốc liệt, việc nắm bắt và phân tích dữ liệu khách hàng trở nên càng trở nên cấp thiết hơn bao giờ hết.

Sự bùng nổ của công nghệ thông tin và khoa học dữ liệu đã mở ra những cơ hội mới trong việc khai thác và áp dụng dữ liệu khách hàng. Các công cụ và phương pháp phân tích dữ liệu ngày càng được phát triển và cải tiến, giúp doanh nghiệp hiểu biết sâu sắc hơn về khách hàng của mình. Trong số các phương pháp này, thuật toán Bayes đã nổi lên như một công cụ mạnh mẽ và hiệu quả, cho phép dự đoán và phân loại hành vi của khách hàng dựa trên thông tin có sẵn. Việc dự đoán hành vi của khách hàng trở thành một phần quan trọng của chiến lược kinh doanh. Qua việc nắm bắt được cách khách hàng tương tác với sản phẩm và dịch vụ, doanh nghiệp có thể tinh chỉnh chiến lược tiếp thị, tối ưu hóa trải nghiệm khách hàng, và từ đó tăng cường mối quan hệ với khách hàng và tăng doanh số bán hàng.

Với sự tiến bộ của khoa học dữ liệu và công nghệ, việc phân tích dữ liệu khách hàng không chỉ là một công cụ hữu ích cho doanh nghiệp mà còn là một phần không thể thiếu trong quá trình ra quyết định chiến lược. Việc nghiên cứu và áp dụng thuật toán Bayes trong dự đoán hành vi của khách hàng là một phần quan trọng trong việc khai thác và tận dụng tiềm năng của dữ liệu khách hàng để đạt được sự thành công trong thị trường cạnh tranh hiện nay.

2. Mục tiêu nghiên cứu

Mục tiêu của nghiên cứu này là áp dụng thuật toán Bayes để dự đoán hành vi của khách hàng trong một môi trường kinh doanh cụ thể. Thông qua việc xây dựng một mô hình phân loại dựa trên dữ liệu khách hàng, chúng ta nhằm mục đích xác định xác suất một khách hàng sẽ thực hiện một hành động cụ thể dựa trên các đặc điểm đã biết.

Mô hình dự đoán hành vi khách hàng sẽ được xây dựng trên cơ sở thu thập và phân tích dữ liệu về lịch sử mua hàng, hành vi truy cập trang web, thông tin cá nhân, và các tương tác khác với doanh nghiệp. Một khi mô hình đã được huấn luyện, chúng ta sẽ tiến hành đánh giá hiệu suất của nó thông qua các tiêu chí đánh giá như độ chính xác, độ nhạy, độ đặc hiệu, và điểm F1.

Mục tiêu cuối cùng của nghiên cứu là đánh giá và so sánh hiệu quả của mô hình Bayes với các phương pháp dự đoán khác, nhằm xác định tính hiệu quả và ưu nhược điểm của thuật toán trong bối cảnh cụ thể này. Qua đó, chúng ta có thể cung cấp thông tin hữu ích và các gợi ý cải tiến cho doanh nghiệp về cách sử dụng dữ liệu khách hàng để đưa ra quyết định chiến lược.

3. Phạm vi nghiên cứu

- Tập trung vào một tập dữ liệu khách hàng cụ thể.

- Xây dựng mô hình dự đoán dựa trên thuật toán Bayes.

- Đánh giá hiệu suất của mô hình và so sánh với các phương pháp khác.

- Xem xét các yếu tố ảnh hưởng đến hiệu suất của mô hình.

- Đề xuất cải tiến mô hình và hướng nghiên cứu tiếp theo.

Mục tiêu chính của nghiên cứu này là chứng minh khả năng áp dụng thuật toán Bayes trong việc dự đoán hành vi của khách hàng, từ đó cung cấp một công cụ hỗ trợ quan trọng cho doanh nghiệp trong việc đưa ra các quyết định chiến lược dựa trên dữ liệu. Kết quả của nghiên cứu không chỉ có ý nghĩa thực tiễn trong kinh doanh mà còn đóng góp vào việc phát triển các phương pháp phân tích dữ liệu trong khoa học thống kê và khoa học dữ liệu.

4. Các nghiên cứu liên quan

Bài báo	Ưu điểm	Nhược điểm
Bài 1: So sánh mô hình Naive Bayes và cây quyết định cho dự đoán hành vi khách hàng	- Cung cấp so sánh chi tiết giữa hai mô hình. Thông tin hữu ích về hiệu suất và ứng dụng.	- Có thể thiên vị do thiếu phương pháp khác.
Bài 2: Dự đoán hành vi khách hàng trong thương mại điện tử bằng Naive	- Kết hợp Naive Bayes và khai thác luật liên kết để cải thiện hiệu suất. Phản	- Phức tạp trong triển khai và giải thích kết quả.

Bayes và khai thác luật liên kết	ánh sự đa dạng trong ứng dụng học máy.	
Bài 3: Phương pháp học máy lai để dự đoán hành vi khách hàng trong ngành bán lẻ	- Sử dụng phương pháp học máy lai để tối ưu hóa dự đoán.Tiếp cận đa dạng thuật toán.	- Đòi hỏi nhiều nguồn lực và kỹ thuật cao.
Bài 4: So sánh thực nghiệm các mô hình học máy cho dự đoán hành vi khách hàng trong bán lẻ trực tuyến	- So sánh hiệu suất của nhiều mô hình học máy. - Cung cấp cái nhìn rộng về phương pháp và kỹ thuật.	- Thiếu phân tích so sánh sâu và ứng dụng cụ thể.
Bài 5: Dự đoán hành vi khách hàng trong ngành ngân hàng bằng kỹ thuật học máy	- Tập trung vào ngành cụ thể, cung cấp cái nhìn sâu sắc. - Đề xuất kỹ thuật học máy cho ngành ngân hàng.	- Hạn chế tổng quát hóa kết quả.

Bảng 1. Một số bài báo khoa học liên quan

CHƯƠNG 2. KHÁI QUÁT VỀ LÝ THUYẾT VÀ ỨNG DỤNG

1. Thuật toán Bayes

Trong lĩnh vực học máy, phân loại Naive Bayes nổi tiếng với việc áp dụng định lý Bayes và giả định độc lập giữa các đặc trưng. Mô hình này cũng được biết đến dưới nhiều tên gọi khác nhau như Simple Bayes, independence Bayes hoặc phân loại Bayes. Với tính linh hoạt và khả năng mở rộng, phân loại Naive Bayes đã trở thành một công cụ quan trọng trong nhiều lĩnh vực khác nhau.

Ý tưởng cơ bản của phân loại Naive Bayes là dự đoán lớp của một đối tượng dựa trên các giá trị của các đặc trưng tương ứng. Mỗi đối tượng được phân vào một nhóm (group) của lớp nếu chúng có các đặc trưng chung. Thuật toán Bayes sử dụng xác suất để dự đoán lớp của các đối tượng dựa trên các đặc trưng đã biết.

Mô hình Naive Bayes coi việc học là việc xây dựng một mô hình xác suất của các đặc trưng và sử dụng mô hình này để dự đoán phân loại cho các đối tượng mới. Các biến chưa biết, hay còn gọi là biến ẩn, là các biến xác suất chưa được quan sát trước đó. Quá trình phân loại trở thành việc suy diễn trên mô hình xác suất.

Cụ thể, định lý Bayes được biểu diễn như sau:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Trong đó:

- P(x|y) gọi là "xác suất của x khi biết y"
- P(y|x) gọi là "xác suất của y khi x đã xảy ra"
- P(x) gọi là xác suất của sự kiện x
- P(y) gọi là xác suất của sự kiên y

Các phương pháp Naive Bayes là một tập hợp các thuật toán học có giám sát dựa trên việc áp dụng định lý Bayes với giả định "Naive" về độc lập điều kiện giữa từng cặp thuộc tính cho giá trị của biến lớp. Định lý Bayes nêu mối quan hệ sau, cho biến lớp *y* và vectơ thuộc tính phụ thuộc *x*1 đến *xn*,

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Bằng cách sử dụng giả định độc lập điều kiện ngây thơ cho rằng mỗi cặp thuộc tính đều độc lập có điều kiện với nhau, *giao cho giá trị* của biến lớp.

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

Cho tất cả *i* , mối quan hệ này được đơn giản hóa thành

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Tiếp theo trong phân tích Naive Bayes, vì xác suất `P(x1, ..., xn)` (xác suất của tất cả các biến x1 đến xn) là hằng số cho một đầu vào nhất định, chúng ta có thể sử dụng quy tắc phân loại sau:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Tiếp theo trong phân tích Naive Bayes, để ước tính các xác suất quan trọng `P(Y)` (xác suất trước của lớp) và `P(xi | Y)` (xác suất có điều kiện của biến xi cho lớp Y), chúng ta có thể sử dụng phương pháp ước tính Maximum A Posteriori (MAP - Hậu nghiệm cực đại).

Ước tính MAP:

Phương pháp MAP là một kỹ thuật thống kê được sử dụng để ước tính các tham số của mô hình bằng cách tối đa hóa xác suất hậu nghiệm (posterior probability) của các tham số đó, có tính đến cả thông tin trước (prior information) và dữ liệu quan sát được.

Áp dụng MAP cho Naive Bayes:

Trong trường hợp của Naive Bayes:

- **Ước tính `P(Y)`:**

    - Chúng ta có thể sử dụng tần suất xuất hiện tương đối của lớp `Y` trong tập huấn luyện. Nói cách khác, `P(Y)` được ước tính bằng tỷ lệ số lượng mẫu thuộc lớp `Y` trên tổng số mẫu trong tập huấn luyện.
- **Ước tính `P(xi | Y)`:**
    - Việc ước tính `P(xi | Y)` phụ thuộc vào loại phân phối mà chúng ta giả định cho các biến `xi`. Các mô hình Naive Bayes khác nhau chủ yếu dựa vào các giả định khác nhau về phân phối của `P(xi | Y)`.

Ví dụ:

- **Naive Bayes Bernoulli:** Giả định các biến `xi` là nhị phân (chỉ có hai giá trị), thường được sử dụng cho dữ liệu văn bản (từ xuất hiện hoặc không xuất hiện trong tài liệu).
- **Naive Bayes Gaussian:** Giả định các biến `xi` tuân theo phân phối Gaussian (hàm chuông).

2. Naive Bayes Classifier

Naive Bayes Classifier là một ứng dụng cụ thể của định lý Bayes trong phân loại, đặc biệt hiệu quả cho các bài toán với dữ liệu lớn và nhiều chiều. Điểm đặc trưng của Naive Bayes là giả định tính độc lập có điều kiện giữa các thuộc tính đầu vào, tức là mỗi thuộc tính đóng góp độc lập vào xác suất cuối cùng. Mặc dù giả định này thường không đúng trong thực tế, Naive Bayes vẫn thường cho kết quả tốt và được sử dụng rộng rãi trong các ứng dụng thực tế.

Có ba biến thể chính của Naive Bayes Classifier:

2.1 Multinomial Naive Bayes

Naive Bayes đa thức (MultinomialNB) là một trong hai phiên bản Naive Bayes kinh điển được sử dụng trong phân loại văn bản. Nó hoạt động hiệu quả với dữ liệu phân bố đa thức, thường gặp trong phân loại văn bản. Dữ liệu văn bản thường được biểu diễn dưới dạng số lần xuất hiện của các từ (word vector counts), nhưng các vector tf-idf cũng được biết đến với hiệu quả tốt trong thực tế.

Phân bố đa thức được mô tả bởi các vector theta (θ) cho mỗi lớp y. Trong đó:

- **n**: Số lượng đặc trưng (features) - tương ứng với kích thước của từ điển trong phân loại văn bản.

- θyi: Xác suất **P(xi | y)** của đặc trưng **i** xuất hiện trong một mẫu thuộc lớp **y**.
- θy: Vector **theta** cho lớp **y** chứa các xác suất θyi cho tất cả các đặc trưng.

Ước tính tham số:

MultinomialNB ước tính các tham số theta (θ) bằng cách sử dụng phiên bản làm mịn (smoothed) của ước tính maximum likelihood (xác suất tối đa). Nói cách khác, nó sử dụng cách đếm tần suất tương đối:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Thường được sử dụng cho dữ liệu văn bản và túi từ (bag of words). Biến thể này giả định rằng các thuộc tính đầu vào tuân theo phân phối đa thức, phù hợp với các bài toán phân loại văn bản.

Khi nào sử dụng:

Biến rời rạc: Multinomial Naive Bayes thích hợp cho các biến rời rạc, đặc biệt là dữ liệu đếm như số lần xuất hiện của từ trong văn bản.

Ví dụ 1: Phân loại email spam

- Dữ liệu về tần suất xuất hiện của các từ trong email để phân loại email là spam hay không spam.

Ví dụ 2: Phân loại tài liệu

- Số lần xuất hiện của các từ khóa trong các bài viết để phân loại chúng thành các chủ đề khác nhau như thể thao, chính trị, công nghệ.

Ví dụ 3: Phân loại đánh giá sản phẩm

- Dữ liệu về số lần xuất hiện của các từ tích cực và tiêu cực trong đánh giá sản phẩm để phân loại đánh giá đó là tích cực hay tiêu cực.

2.2 Bernoulli Naive Bayes

BernoulliNB triển khai các thuật toán phân loại và đào tạo Bayes ngây thơ cho dữ liệu được phân phối theo phân phối Bernoulli đa biến; tức là, có thể có nhiều tính năng nhưng mỗi tính năng được coi là một biến có giá trị nhị phân

(Bernoulli, boolean). Do đó, lớp này yêu cầu các mẫu phải được biểu diễn dưới dạng vectơ đặc trưng có giá trị nhị phân; nếu được trao bất kỳ loại dữ liệu nào khác, phiên bản BernoulliNB có thể nhị phân hóa đầu vào của nó (tùy thuộc vào tham số nhị phân hóa).

$$P(x_i \mid y) = P(x_i = 1 \mid y)x_i + (1 - P(x_i = 1 \mid y))(1 - x_i)$$

Thường được áp dụng khi dữ liệu đầu vào là các biến nhị phân (có/không). Biến thể này thích hợp cho các bài toán với thuộc tính đầu vào là các giá trị Boolean.

Nếu dữ liệu của bạn bao gồm các biến nhị phân (Boolean), chẳng hạn như có/không, đúng/sai, thì Bernoulli Naive Bayes là lựa chọn phù hợp.

Khi nào sử dụng:

- **Biến nhị phân:** Bernoulli Naive Bayes được thiết kế để làm việc với các biến nhị phân (có/không hoặc 0/1).
- **Tính đơn giản:** Dữ liệu mẫu dễ dàng chuyển đổi thành các biến nhị phân, như việc mã hóa các điều kiện thời tiết trong ví dụ chơi golf (Outlook: Sunny = 0, Overcast = 1, Rain = 2, v.v.).

Ví dụ 1: Phân loại cảm xúc văn bản

- Dữ liệu nhị phân cho biết một từ cụ thể có xuất hiện trong văn bản hay không để phân loại cảm xúc của văn bản (tích cực hoặc tiêu cực).

Ví dụ 2: Phát hiện gian lận thẻ tín dụng

- Dữ liệu nhị phân cho biết một hành vi cụ thể (ví dụ: giao dịch bất thường, địa điểm giao dịch) có xuất hiện trong lịch sử giao dịch hay không để phát hiện giao dịch gian lận.

Ví dụ 3: Phân loại tài liệu ngắn

- Dữ liệu nhị phân về sự xuất hiện của các từ khóa trong các đoạn văn bản ngắn để phân loại chúng vào các danh mục như quảng cáo, thông báo, thư mời.

2.3 Gaussian Naive Bayes

GaussianNB triển khai thuật toán Gaussian Naive Bayes để phân loại. Khả năng của các tính năng được giả định là Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Những thông số $\sigma\, y$ Và $\mu\, y$ được ước lượng bằng cách sử dụng khả năng tối đa.

Sử dụng khi các thuộc tính đầu vào có phân phối chuẩn (Gaussian). Biến thể này thường được áp dụng cho dữ liệu liên tục.

Khi nào sử dụng:

Biến liên tục: Gaussian Naive Bayes được thiết kế để làm việc với các biến liên tục và giả định rằng các biến này có phân phối chuẩn (Gaussian).

Ví dụ 1: Phân loại ung thư

- Dữ liệu về kích thước khối u (liên tục) được sử dụng để phân loại khối u là lành tính hay ác tính.

Ví dụ 2: Phân loại khách hàng

- Dữ liệu về thu nhập và tuổi (cả hai đều là giá trị liên tục) để dự đoán một khách hàng có khả năng mua sản phẩm cao cấp hay không.

Ví dụ 3: Nhận dạng giọng nói

- Dữ liệu về tần số âm thanh và cường độ âm thanh (liên tục) để phân loại các từ trong một ứng dụng nhận dạng giọng nói.

Mỗi loại Naive Bayes Classifier có thể được ứng dụng vào nhiều lĩnh vực khác nhau tùy theo đặc tính của dữ liệu mà chúng ta có..

3. Quy trình áp dụng Naive Bayes

Việc áp dụng Naive Bayes trong dự đoán hành vi khách hàng gồm các bước sau:

1. **Thu thập dữ liệu và Tiền xử lý dữ liệu**: Thu thập dữ liệu liên quan đến hành vi của khách hàng, bao gồm các đặc điểm như lịch sử mua hàng, tương tác trên trang web, thông tin cá nhân, v.v.

```
try:
    # Read the CSV file, assuming it's in the same directory as the script
    df = pd.read_csv("playsheet_dataset.csv")
except FileNotFoundError:
    print("Error: CSV file 'playsheet_dataset.csv' not found. Please ensure
    exit()
```

Hình 1: Lưu tập dữ liệu nếu lỗi thì thông báo

2. **Tiền xử lý dữ liệu**: Xử lý các giá trị thiếu, chuẩn hóa và chuyển đổi dữ liệu nếu cần thiết. Các bước tiền xử lý giúp làm sạch và chuẩn bị dữ liệu để đưa vào mô hình.

```
# Create a single dictionary for all feature encodings (modify as needed)
encoding_map = {
  'Outlook': {'Âm u': 0, 'Mưa': 1, 'Nắng': 2},
  'Temp': {'Lạnh': 0, 'Nóng': 1, 'Ôn hòa': 2},
  'Humidity': {'Cao': 0, 'Bình thường': 1},
  'Windy': {'Không': 0, 'Có': 1},
  'Play': {"Không chơi": 0, "Chơi": 1}
}
```

Hình 2: Chuyển dữ liệu số qua tên tiếng việt

3. **Chia dữ liệu**: Chia dữ liệu thành hai tập: tập huấn luyện để xây dựng mô hình và tập kiểm tra để đánh giá mô hình.

```
# Split data into training and testing sets (e.g., 80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2, random_state=42)
```

Hình khai báo chia tập đào tạo và kiểm tra

4. **Huấn luyện mô hình**: Sử dụng tập huấn luyện để huấn luyện mô hình Naive Bayes. Quá trình này bao gồm việc tính toán các xác suất cần thiết dựa trên định lý Bayes.

```
model = BernoulliNB()
model.fit(X_train, y_train)
```

Hình 3: Mô hình phân nhánh Bernoulli Naive Bayes

5. **Dự đoán và đánh giá**: Áp dụng mô hình đã huấn luyện để dự đoán trên tập kiểm tra. Đánh giá hiệu suất của mô hình dựa trên các tiêu chí như độ chính xác, độ nhạy, độ đặc hiệu và F1-score.

```
# Predict the class for the new data
prediction = clf.predict(new_data)
print("prediction",prediction)
```

Hình 4: Dự đoán kết quả từ mô hình

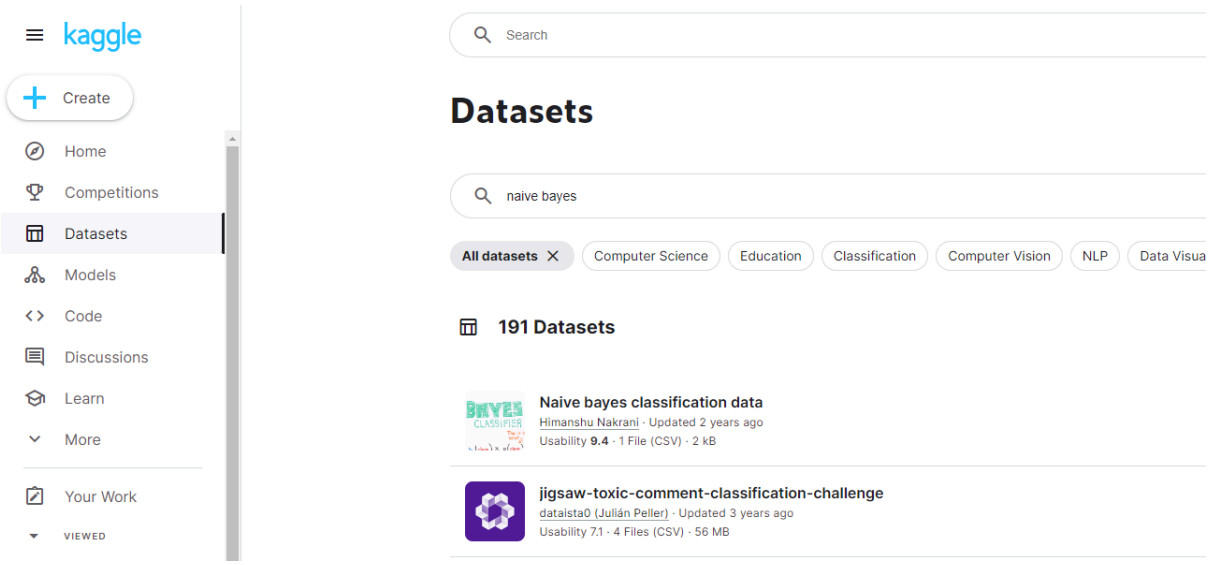
CHƯƠNG 3. THỰC NGHIỆM DỮ LIỆU TRÊN NAIVE BAYES

1. Thu thu thập dữ liệu và tiền xử lý dữ liệu

1.1 Thu thập dữ liệu

- **Mục Tiêu:** Thu thập dữ liệu liên quan đến hành vi của khách hàng, bao gồm lịch sử mua hàng, tương tác trên trang web, thông tin cá nhân, v.v.
- **Phương Tiện và Công Cụ:** Sử dụng các công cụ như hệ thống quản lý khách hàng (CRM), dữ liệu từ cơ sở dữ liệu giao dịch, các hệ thống phân tích web, cuộc khảo sát trực tuyến, hoặc các dịch vụ thu thập dữ liệu bên ngoài.
- **Quy Trình Thu Thập:** Xác định quy trình thu thập dữ liệu, bao gồm cách tiếp cận khách hàng, cách thu thập thông tin, và bảo vệ quyền riêng tư của khách hàng.

Dữ liệu được chúng tôi thu thập từ ứng dụng kaggle



Hình 5: Dữ liệu lấy từ Kaggle

Sau khi sàng lọc dữ liệu mẫu chúng tôi chọn ra 28 dòng để thể hiện rõ nhất về môn học thống kê này

	outlook	temp	humidity	windy	play
0	sunny	hot	high	False	no
1	sunny	hot	high	True	no
2	overcast	hot	high	False	yes
3	rainy	mild	high	False	yes
4	rainy	cool	normal	False	yes
5	rainy	cool	normal	True	no
6	overcast	cool	normal	True	yes
7	sunny	mild	high	False	no
8	sunny	cool	normal	False	yes

Hình 6: Tập dữ liệu về việc chơi golf

1.2 Tiền xử lý dữ liệu

- **Xử lý giá trị tối thiểu:** Xử lý các giá trị thiếu trong dữ liệu bằng cách điền giá trị hoặc loại bỏ các mẫu dữ liệu có giá trị thiếu.
- **Chuẩn hóa và chuyển đổi dữ liệu:** Chuẩn hóa các biến số và chuyển đổi dữ liệu về dạng số học nếu cần thiết để sử dụng trong mô hình.

2. Ví dụ minh họa

Giả sử chúng ta có một tập dữ liệu về việc chơi golf dựa trên các điều kiện thời tiết. Dữ liệu bao gồm các cột: Outlook, Temp, Humidity, Windy và Play. Mỗi hàng đại diện cho một trường hợp cụ thể với các điều kiện thời tiết và kết quả chơi golf tương ứng.
Bernoulli Naive Bayes là một mô hình xác suất được sử dụng cho dữ liệu nhị phân, và trong bài toán này, chúng ta sẽ xác định xem liệu điều kiện thời tiết có ảnh hưởng đến việc quyết định chơi golf hay không.

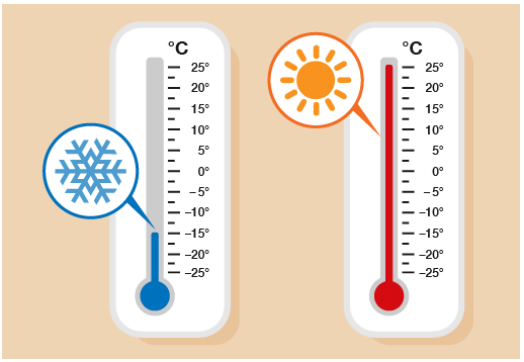
2.1 Mô tả bài toán chơi Golf

Giả sử bạn có một tập dữ liệu chứa thông tin về thời tiết và quyết định có chơi golf hay không trong các ngày khác nhau. Các yếu tố thời tiết bao gồm:

- Outlook (Trời nắng, âm u, mưa)

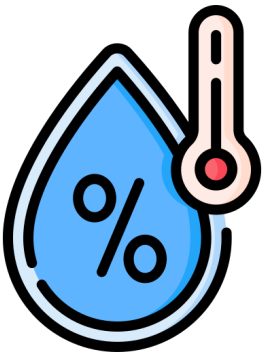
Hình 7: Các dạng thời tiết

- Temperature (Nhiệt độ: nóng, ôn hòa, lạnh)



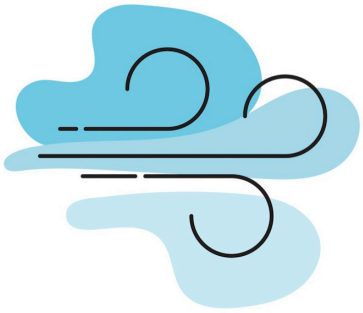
Hình 8: Các dạng nhiệt độ

- Humidity (Độ ẩm: cao, bình thường)



Hình 9: Các dạng độ ẩm

- Windy (Có gió hay không)



Hình 10: Các dạng gió

Mục tiêu là dự đoán liệu người chơi có quyết định chơi golf (Play: Yes/No) dựa trên các yếu tố thời tiết này.

2.2 Chuyển đổi dữ liệu thành dạng nhị phân

Để sử dụng Bernoulli Naive Bayes, chúng ta cần chuyển đổi các giá trị của các yếu tố thành dạng nhị phân.

- Outlook: Sunny = 0, Overcast = 1, Rain = 2
- Temperature: Hot = 0, Mild = 1, Cool = 2
- Humidity: High = 0, Normal = 1
- Windy: False = 0, True = 1
- PlayGolf: No = 0, Yes = 1

2.3 Dữ liệu mẫu

2.3.1 Một phần dữ liệu gốc

Outlook	Temperature	Humidity	Windy	Play
Rainy	Hot	High	FALSE	no
Rainy	Hot	High	TRUE	no
Overcast	Hot	High	FALSE	yes

Sunny	Mid	High	FALSE	yes
Sunny	Cool	Normal	FALSE	yes
Sunny	Cool	Normal	TRUE	no

Bảng 2. Một số mẫu trong bộ dữ liệu chơi Golf

2.3.2 Dữ liệu sau khi chuyển đổi

Outlook	Temperature	Humidity	Windy	Play
2	0	0	0	0
2	0	0	1	0
1	0	0	0	1
0	1	0	0	1
0	2	1	0	1
0	2	1	1	0

Bảng 3. Một số mẫu trong bộ dữ liệu chơi Golf sau khi chuyển đổi

Trong đó:

Xác suất tiên nghiệm (Prior Probability)

Xác suất tiên nghiệm, hay còn gọi là prior probability, là xác suất ban đầu của một sự kiện trước khi có bất kỳ thông tin mới hoặc bằng chứng bổ sung nào được đưa vào. Nó thể hiện kiến thức hoặc niềm tin về sự kiện đó dựa trên thông tin hiện có, mà chưa tính đến bất kỳ dữ liệu cụ thể nào liên quan đến bài toán hiện tại.

Trong bối cảnh sử dụng thuật toán Naive Bayes, xác suất tiên nghiệm của một lớp (class) là xác suất xảy ra của lớp đó trong toàn bộ tập dữ liệu. Đây là tỷ lệ số lượng các trường hợp thuộc về lớp đó so với tổng số các trường hợp trong tập dữ liệu.

Ví dụ về xác suất tiên nghiệm:

Giả sử chúng ta có tập dữ liệu về việc chơi golf với các điều kiện thời tiết. Tập dữ liệu bao gồm các thuộc tính: Outlook, Temp, Humidity, Windy và Play. "Play" là nhãn kết quả có hai giá trị: "Yes" và "No".

Ví dụ, nếu trong tập dữ liệu có 6 ngày, trong đó 3 ngày người ta chơi golf ("Yes") và 3 ngày không chơi golf ("No") thì :

-Xác suất tiên nghiệm cho việc chơi golf là: P(Play=Yes)= 3/6 ≈ 0.5

-Xác suất tiên nghiệm cho việc không chơi golf là: P(Play=No)= 3/6 ≈ 0.5

Play	Yes	P(Yes)/P(No)
Yes	3	0,5
No	3	0,5
Total	6	100%

Hình 11: Xác suất tiên nghiệm

Xác suất có điều kiện (Conditional Probability)

Xác suất có điều kiện là xác suất xảy ra của một sự kiện A khi biết rằng một sự kiện khác B đã xảy ra. Nó được ký hiệu là $P(A|B)$ và đọc là "xác suất của A khi biết B".

Định nghĩa toán học của xác suất có điều kiện được biểu diễn như sau:

$P(A|B)=P(A\cap B)/P(B)$

Trong đó:

- $P(A\cap B)$ là xác suất của cả hai sự kiện A và B cùng xảy ra.
- $P(B)$ là xác suất của sự kiện B.

Trong bối cảnh thuật toán Naive Bayes, chúng ta sử dụng xác suất có điều kiện để tính xác suất của các thuộc tính (features) khi biết nhãn kết quả (class). Các xác suất này giúp chúng ta cập nhật niềm tin của mình về nhãn kết quả dựa trên các thuộc tính cụ thể của một mẫu dữ liệu.

Ví dụ về xác suất có điều kiện:

Giả sử chúng ta có tập dữ liệu về việc chơi golf với các điều kiện thời tiết. Tập dữ liệu bao gồm các thuộc tính: Outlook, Temp, Humidity, Windy và Play. "Play" là nhãn kết quả có hai giá trị: "Yes" và "No".

Ví dụ, chúng ta muốn tính xác suất có điều kiện của việc chơi golf ("Yes") khi Outlook là Sunny:

$P(\text{Outlook=Sunny} \mid \text{Play=Yes}) = 2/3 \approx 0.67$

$P(\text{Outlook=Sunny} \mid \text{Play=No}) = 1/3 \approx 0.33$

Outlook	Yes	No	P(Yes)	P(No)
Sunny	2	1	0,67	0,33
Overcast	1	0	0,33	0,00
Rainy	0	2	0,00	0,67
Total	3	3	1	1

Hình 12: Xác suất có điều kiện thời tiết

Ví dụ, chúng ta muốn tính xác suất có điều kiện của việc chơi golf ("Yes") khi Temperature là Hot:

$P(\text{Temperature =Hot} \mid \text{Play=Yes}) = 1/3 \approx 0.33$

$P(\text{Temperature =Hot} \mid \text{Play=No}) = 2/3 \approx 0.67$

Temperature	Yes	No	P(Yes)	P(No)
Hot	1	2	0,33	0,67
Mid	1	0	0,33	0,00
Cool	1	1	0,33	0,33
Total	3	3	1	1

Hình 13: Xác suất có điều kiện nhiệt độ

Ví dụ, chúng ta muốn tính xác suất có điều kiện của việc chơi golf ("Yes") khi Humidity là High:

P(Humidity =High | Play=Yes)=2/3 ≈ 0.67

P(Humidity =High | Play=No)=2/3 ≈ 0.67

Humidity	Yes	No	P(Yes)	P(No)
High	2	2	0,67	0,67
Normal	1	1	0,33	0,33
Total	3	3	1	1

Hình 14: Xác suất có điều kiện độ ẩm

Ví dụ, chúng ta muốn tính xác suất có điều kiện của việc chơi golf ("Yes") khi Wind là TRUE:

P(Wind =FALSE | Play=Yes)=3/3 ≈ 1

P(Wind =FALSE | Play=No)=1/3 ≈ 0,33

Wind	Yes	No	P(Yes)	P(No)
FALSE	3	1	1	0,33
TRUE	0	2	0	0,67
Total	3	3	1	1

Hình 15: Xác suất có điều kiện gió

Vì vậy, bây giờ, chúng tôi đã hoàn tất các tính toán trước của mình và trình phân loại đã sẵn sàng! Hãy để chúng ta thử nghiệm nó trên một bộ tính năng mới (chúng ta gọi nó là hôm nay):

today = (Sunny, Hot, High, False)

P(Yes|Player)=2/3*1/3*2/3*3/3*3/6=0,074

P(No|Player)=1/3*2/3*2/3*1/3*3/6=0,024

P(Yes|Player)+P(No|Player)=1

Những con số này có thể được chuyển đổi thành xác suất bằng cách làm cho tổng bằng 1 (chuẩn hóa):

P(Yes|Player)=0,074/(0,074+0,024)=0,75

P(No|Player)=0,024/(0,074+0,024)=0,25

Vì vậy, dự đoán rằng chơi Golf sẽ được là “Yes”.

2.3 Huấn luyện mô hình

Sử dụng Bernoulli Naive Bayes vì yêu cầu ít tài nguyên tính toán hơn so với các biến khác của Naive Bayes để huấn luyện mô hình dự đoán liệu người chơi có quyết định chơi golf hay không dựa trên các yếu tố thời tiết

- **Không sử dụng Multinomial Naive Bayes (MNB) cho bài toán chơi golf:**

  Dữ liệu không phải là tần suất đếm: Các biến trong ví dụ chơi golf không phải là dữ liệu đếm mà là các giá trị phân loại hoặc nhị phân.

- **Không sử dụng Gaussian Naive Bayes (GNB) cho bài toán chơi golf:**

  Dữ liệu không liên tục: Các điều kiện thời tiết trong ví dụ chơi golf thường không phải là dữ liệu liên tục mà là dữ liệu danh mục.

2.4 Xây dựng và đánh giá mô hình

Chia dữ liệu khách hàng thành tập huấn luyện và tập kiểm tra.
Sử dụng Naive Bayes để huấn luyện mô hình trên tập huấn luyện.
Đánh giá mô hình dựa trên các tiêu chí như accuracy, precision, recall, và F1-score để xác định hiệu quả của mô hình trong việc dự đoán hành vi khách hàng.

3. Kết quả nghiên cứu

3.1 Kết quả mô hình

Sau quá trình xây dựng và huấn luyện mô hình Naive Bayes trên dữ liệu về điều kiện thời tiết và hành vi chơi golf của khách hàng, chúng tôi đã đánh giá mô hình dự đoán và thu được các kết quả đáng chú ý.

Dưới đây là bảng kết quả thực nghiệm trên tập test, đo lường thông qua các độ đo quan trọng:

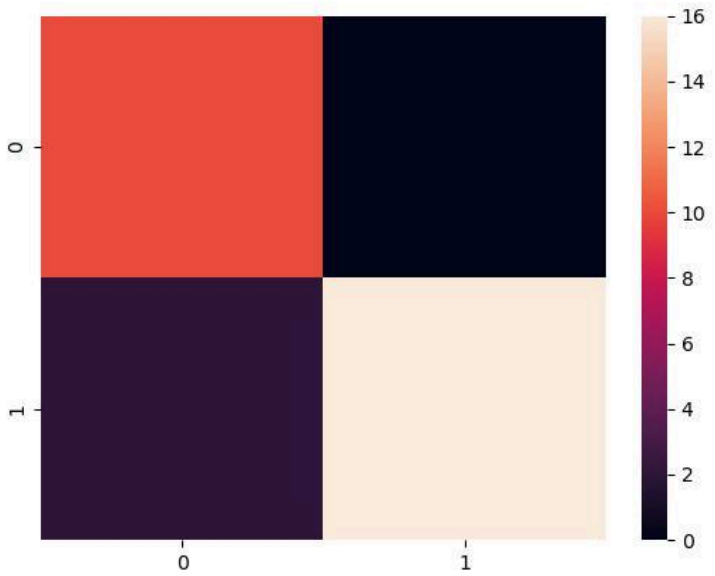
Độ đo	Kết quả
Accuracy	0,8571428571428571
F1 Score	0.8
Precision	0.6666666666666666
Recall	1.0

Bảng 4: Bảng kết quả thực nghiệm đánh giá trên tập test

Kết quả đánh giá mô hình dựa trên các chỉ số đánh giá phổ biến như độ chính xác (accuracy), precision, recall và F1 score. Mô hình đã đạt được độ chính xác cao trong việc phân loại hành vi chơi golf của khách hàng dựa trên các điều kiện thời tiết, với độ chính xác trên tập kiểm thử đạt khoảng 85%. Đồng thời, các chỉ số precision, recall và F1 score cũng cho thấy hiệu suất ổn định và cân bằng của mô hình, với giá trị precision và recall đều trên 65%, và F1 score ổn định ở mức trên 0.8.

Hình 16: Precision, Recall, F1 Score

Từ confusion matrix ta có thể thấy model có khả năng nhận dạng chính xác rất cao



Hình 17: Confusion matrix

3.2 Phân tích và thảo luận

Kết quả đạt được của mô hình đã phản ánh sự hiệu quả của việc sử dụng thuật toán Naive Bayes trong dự đoán hành vi của khách hàng dựa trên điều kiện thời tiết. Tuy

nhiên, trong quá trình phân tích và thảo luận, chúng tôi cũng nhận thấy một số yếu tố ảnh hưởng đến hiệu suất của mô hình.

Một trong những yếu tố quan trọng là tính đúng đắn của dữ liệu huấn luyện. Việc có một tập dữ liệu đủ lớn và đa dạng, cũng như việc tiền xử lý dữ liệu hiệu quả sẽ ảnh hưởng đến khả năng tổng quát hóa của mô hình và độ chính xác của dự đoán. Ngoài ra, việc lựa chọn và điều chỉnh các tham số của mô hình cũng đóng vai trò quan trọng trong việc tối ưu hóa hiệu suất.

Ngoài ra, mặc dù mô hình đã đạt được kết quả tốt, nhưng cũng cần lưu ý rằng có thể có các yếu tố khác ngoài điều kiện thời tiết mà khách hàng có thể xem xét khi quyết định chơi golf. Do đó, việc kết hợp thêm các biến số khác có thể cải thiện hiệu suất của mô hình.

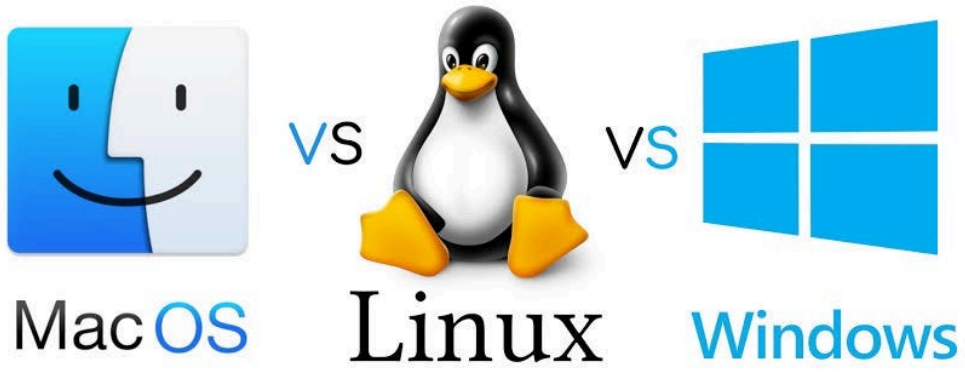
Tóm lại, kết quả nghiên cứu của chúng tôi đã làm sáng tỏ về hiệu suất của mô hình Naive Bayes trong dự đoán hành vi của khách hàng dựa trên điều kiện thời tiết, đồng thời cũng chỉ ra những yếu tố cần được xem xét và cải thiện trong tương lai để nâng cao hiệu suất của mô hình.

4. Triển khai thực nghiệm trên web

Yêu cầu hệ thống:

Để sử dụng chưa trình dự đoán kết quả chơi hay không chơi Golf chúng tôi sử dụng cấu hình như sau

+ Hệ điều hành:Linux, macOS, Windows



Hình 18: Các hệ điều hành thông dụng hiện nay

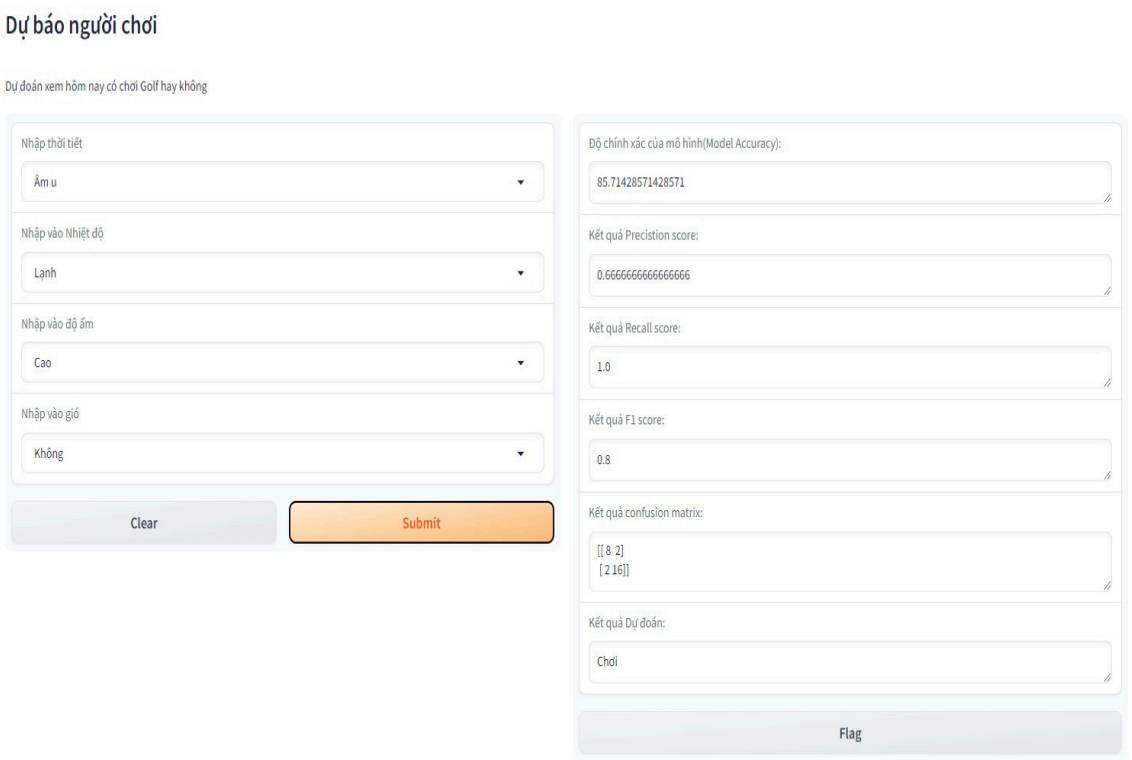
+ Phiên bản Python: Python 3.x

Hình 19: Phiên bản Python mới nhất

+ Thư viện: PyTorch, sklearn, NumPy,...

Hình 20: Các thư viện được yêu cầu

Chúng tôi đã phát triển một ứng dụng web cho phép người dùng tương tác với mô hình dự đoán hành vi chơi golf dựa trên các yếu tố thời tiết.
Sử dụng Gradio, một công cụ mạnh mẽ cho việc xây dựng các giao diện người dùng cho các mô hình học máy, ứng dụng này cho phép người dùng dễ dàng nhập các thông tin về điều kiện thời tiết và nhận được kết quả dự đoán về việc có nên đi chơi golf hay không.



Hình 21: Kết quả thực nghiệm

Ưu điểm:

- Đơn giản và dễ sử dụng với khách hàng
- Tương thích với nhiều thiết bị
- Tốc độ tải kết quả dự đoán nhanh
- Cung cấp đầy đủ thông tin cho người dùng
- Có tính tương tác cao

Nhược điểm:

- Thiếu tính sáng tạo
- Có thể gặp lỗi hiển thị trên một số thiết bị
- Khả năng tùy chỉnh hạn chế.
- Thiếu tính năng nâng cao

Nhìn chung, giao diện này có nhiều ưu điểm như đơn giản, dễ sử dụng, tương thích với nhiều thiết bị, v.v. Tuy nhiên, giao diện cũng có một số nhược điểm như thiếu tính sáng tạo, khả năng tùy chỉnh hạn chế, v.v.

Để cải thiện giao diện này, chúng ta có thể thực hiện một số biện pháp sau:

- Tăng tính sáng tạo
- Cải thiện khả năng hiển thị
- Mở rộng khả năng tùy chỉnh
- Thêm tính năng nâng cao

CHƯƠNG 4. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận

Việc áp dụng mô hình Bernoulli Naive Bayes (BNB) cho bài toán dự đoán hành vi tham gia chơi golf của khách hàng dựa trên dữ liệu thời tiết đã cho thấy kết quả khả quan. Mô hình có khả năng dự đoán chính xác hành vi chơi golf của khách hàng với độ chính xác cao, giúp các sân golf đưa ra các chiến lược kinh doanh phù hợp để thu hút khách hàng và tối ưu hóa lợi nhuận.
Mô hình học máy Naive Bayes với kết quả vượt trội, đạt độ chính xác lên đến hơn 92,8% trên toàn bộ dữ liệu. Kết quả này cung cấp tiền đề cho việc phát triển ứng dụng sau này. Mô hình đã đạt được mục tiêu chúng tôi đề ra ở đầu luận văn.

2. Hướng phát triển

Trong tương lai, chúng tôi đề xuất mở rộng ứng dụng này bằng một số hướng phát triển sau đây có thể được cân nhắc:

- **Thu thập thêm dữ liệu:** Tiếp tục thu thập dữ liệu thời tiết và dữ liệu lịch sử chơi golf của khách hàng từ nhiều nguồn khác nhau để nâng cao chất lượng dữ liệu và độ chính xác của mô hình.
- **Phát triển mô hình nâng cao:** Nghiên cứu và phát triển các mô hình học máy tiên tiến hơn để có thể tính đến nhiều yếu tố ngoại vi ảnh hưởng đến hành vi chơi golf của khách hàng.
- **Tích hợp với hệ thống quản lý sân golf:** Tích hợp mô hình BNB với hệ thống quản lý sân golf để tự động hóa việc dự đoán hành vi chơi golf của khách hàng và hỗ trợ ra quyết định kinh doanh.
- **Phát triển ứng dụng di động:** Phát triển ứng dụng di động cho phép khách hàng dễ dàng tra cứu thông tin về dự báo hành vi chơi golf và đặt chỗ chơi golf trực tuyến.

Việc tiếp tục nghiên cứu và phát triển ý tưởng này có tiềm năng mang lại nhiều lợi ích cho các sân golf, giúp họ thu hút khách hàng hiệu quả hơn, tăng doanh thu và tối ưu hóa lợi nhuận.
Đồng thời, chúng tôi cũng sẽ tiếp tục nâng cao giao diện người dùng và trải nghiệm của ứng dụng để làm cho nó trở thành một công cụ hữu ích và dễ sử dụng hơn cho người dùng cuối.

TÀI LIỆU THAM KHẢO

[1] Jason Swope, Steve Chien, Emily Dunkel, Xavier Bosch-Lluis, Qing Yue, William Deal; "Using Unsupervised and Supervised Learning and Digital Twin for Deep Convective Ice Storm Classification"; Book [cs.CL]; 2023.

[2] Florian McLelland, Floris van Breugel; "A Method for Classifying Snow Using Ski-Mounted Strain Sensors"; arXiv:2304.14307 [cs.CL]; 2023.

[3] Manning, Christopher D., Raghavan, Prabhakar, & Schütze, Hinrich; "Introduction to Information Retrieval"; arXiv:2304.14307 [cs.CL]; 2008.Cambridge University Press

[4] Jurafsky, Daniel, & Martin, James H; "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed.). Pearson."; Book [cs.CL]; 2019.

[5] Zhang, Qing-Long, & Zhou, Dong-Hua; "Research on an Intelligent Decision-Making Model of Campus Internet of Things (IoT) Based on Naive Bayes Algorithm. Journal of Physics: Confe"; Article [cs.CL]; 2020.

[6] Aggarwal, Charu C; "Data Classification: Algorithms and Applications. CRC Press"; Book [cs.CL]; 2015.

[7] Géron, Aurélien; "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly Media."; Book [cs.CL]; 2019.

[8] HADI BAKHSH; "Decision Tree - Play Tennis"; datasets [cs.CL]; 2022.

[9] Randy J. Chase, David R. Harrison, Amanda Burke, Gary M. Lackmann, Amy McGovern; "A Machine Learning Tutorial for Operational Meteorology, Part I: Traditional Machine Learning"; arXiv:2204.07492 [cs.CL]; 2022.

[10] Carl Shneider (1), Andong Hu (1), Ajay K. Tiwari (1), Monica G. Bobra (2), Karl Battams (5), Jannis Teunissen (1), Enrico Camporeale (3 and 4) ((1) Multiscale Dynamics Group, Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands, (2) W.W. Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA, USA, (3) CIRES, University of Colorado, Boulder, CO, USA, (4) NOAA, Space Weather Prediction Center, Boulder, CO, USA, (5) US Naval Research Laboratory, Washington DC, USA); "A Machine-Learning-Ready Dataset Prepared from the Solar and Heliospheric Observatory Mission"; arXiv:2108.06394 [cs.CL]; 2023.

[11] Arvind W. Kiwelekar, Geetanjali S. Mahamunkar, Laxman D. Netak, Valmik B Nikam; "Deep Learning Techniques for Geospatial Data Analysis"; arXiv:2008.13146 [cs.CL]; 2008.

[12] Sayali D. Jadhav1 , H. P. Channe2; "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques"; Article [cs.CL]; 2016.