Dự Đoán Khả Năng Mưa

*Dựa trên 3 phương pháp Naive Bayes, Markov Chains và Monte Carlo

1st Lương Anh Huy
Trường Đại học Công Nghệ Thông Tin
Thành phố Hồ Chí Minh, Việt Nam
22520550@gm.uit.edu.vn

2nd Phạm Đông Hưng
Trường Đại học Công Nghệ Thông Tin
Thành phố Hồ Chí Minh, Việt Nam
22520521@gm.uit.edu.vn

3rd Phan Công Minh
Trường Đại học Công Nghệ Thông Tin
Thành phố Hồ Chí Minh, Việt Nam
22520884@gm.uit.edu.vn

4th Hồng Khải Nguyên
Trường Đại học Công Nghệ Thông Tin
Thành phố Hồ Chí Minh, Việt Nam
22520967@gm.uit.edu.vn

I. Giới thiệu

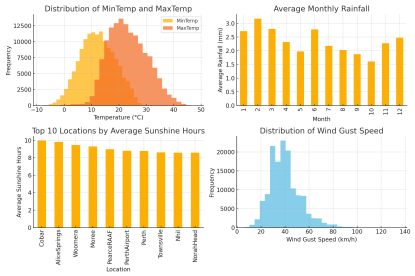
Dự đoán khả năng mưa là một trong những thách thức lớn trong khí tượng học, ảnh hưởng trực tiếp đến các lĩnh vực như nông nghiệp, quản lý nguồn nước và phòng chống thiên tai.

Các phương pháp toán học như Naive Bayes, Markov Chains và Monte Carlo đã được ứng dụng để giải quyết bài toán này. Báo cáo này sẽ trình bày các phương pháp trên, so sánh kết quả dự đoán của chúng và đưa ra những nhận xét về hiệu quả của mỗi phương pháp trong việc dự đoán khả năng mưa.

II. Dataset

Dữ liệu bao gồm các thông tin liên quan đến thời tiết với các cột như ngày tháng, địa điểm, nhiệt độ, lượng mưa, sự bốc hơi, số giờ nắng và nhiều thông số khác như hướng và tốc độ gió, áp suất không khí, độ mây vào buổi sáng và chiều, nhiệt độ vào buổi sáng và chiều, cũng như tình trạng mưa hôm nay và ngày mai.

Biểu đồ tóm tắt các đặc điểm chính của dữ liệu:



Phía trên gồm các biểu đồ:

- **Phân bố Nhiệt độ Tối thiểu và Tối đa:** Biểu đồ cho thấy tần suất xuất hiện của nhiệt độ tối thiểu và tối đa, giúp hình dung khoảng nhiệt độ phổ biến trong dữ liệu.
- **Lượng Mưa Trung Bình Hàng Tháng:** Lượng mưa trung bình mỗi tháng cho thấy xu hướng mưa qua các tháng, hỗ trợ phân tích về mùa mưa.
- **Số Giờ Nắng Trung Bình Theo Địa Điểm:** Biểu đồ top 10 địa điểm có số giờ nắng trung bình cao nhất, cho thấy sự khác biệt về độ nắng giữa các vùng.
- **Phân bố Tốc Độ Gió:** Phân bố tốc độ gió cho biết mức độ phổ biến của các tốc độ gió mạnh, đặc biệt là tốc độ gió giật.

III. Cơ sở lý thuyết

- **Naive Bayes:** Naive Bayes dựa trên định lý Bayes, sử dụng dữ liệu lịch sử và các biến độc lập như nhiệt độ, độ ẩm để tính toán xác suất khả năng mưa.
- **Markov Chains:** Phương pháp chuỗi Markov dựa trên nguyên lý rằng trạng thái hiện tại phụ thuộc vào trạng thái trước đó. Trong việc dự đoán khả năng mưa, chuỗi Markov có thể sử dụng dữ liệu lịch sử về sự xuất hiện hay không của mưa trong các ngày trước đó để ước tính xác suất mưa trong ngày tiếp theo.
- **Monte Carlo:** Phương pháp Monte Carlo là một kỹ thuật mô phỏng ngẫu nhiên, dựa vào việc tạo ra nhiều kịch bản dự đoán khác nhau để tính toán xác suất khả năng mưa.

A. Naive Bayes

Naive Bayes là một trong nhóm các thuật toán áp dụng định lý Bayes với một giả định khá ngây thơ (Naive), giả định cho rằng các thuộc tính (biến) có độ quan trọng như nhau và các thuộc tính (biến) độc lập có điều kiện khi được cho lớp/nhãn.

Định lý Bayes:

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

Naive Bayes là một thuật toán phân loại cho các vấn đề phân loại nhị phân và đa lớp. Kỹ thuật này dễ hiểu nhất khi được mô tả bằng các giá trị đầu vào nhị phân hoặc phân loại.

Thuật toán Naive Bayes tính xác suất cho các yếu tố, sau đó chọn kết quả với xác suất cao nhất. Tuy nhiên, ta cần lưu ý giả định của thuật toán Naive Bayes là các yếu tố đầu vào được cho là độc lập với nhau.

Luật Bayes

Học phân lớp khi có dữ liệu đến

- Evidence X = dữ liệu.
- Event Y = giá trị lớp của dữ liệu.

Naive:

$$P(Y|X) = \frac{P(X_1 \mid Y)P(X_2 \mid Y)\dots P(X_n \mid Y)P(Y)}{P(X)}$$

B. Markov Chains

Markov Chain (chuỗi Markov) là một mô hình hay tiến trình ngẫu nhiên mô tả một chuỗi các sự kiện có khả năng xảy ra, mà xác suất xảy ra sự kiện tiếp theo phụ thuộc chỉ vào sự kiện hiện tại, nghĩa là các sự kiện xảy ra trong quá khứ sẽ không được ghi nhớ, sự kiện trong tương lai chỉ phụ thuộc sự kiện hiện tại của mô hình.

Ví dụ, ta định nghĩa Markov Chain có:

- Tập trạng thái là $Q = q_1, q_2, q_3$
- Ma trận chuyển đổi giữa các trạng thái là $P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$

Markov Chain sẽ có hình dạng trông giống như một đồ thị có hướng, trọng số với các node trên đồ thị là các trạng thái trong tập trạng thái của Markov Chain, các trọng số trên các cạnh mô tả xác suất chuyển p_{ij} là xác suất để di chuyển từ trạng thái q_i đến trạng thái q_j. Nếu $p_{ij} = 0$, ta ngầm hiểu rằng không thể di chuyển giữa hai trạng thái q_i và q_j.

Các xác suất chuyển một bước sẽ có các tính chất sau:

$$p_{ij} \geq 0; \quad \sum_{j \in \mathbb{Z}} p_{ij} = 1; \quad i, j \in \mathbb{Z}.$$

Xác suất sau m-bước chuyển của xích Markov ký hiệu bởi:

$$p_{ij}^{(m)} = P(X_{n+m} = j \mid X_n = i); \quad m = 1, 2, \dots$$

Ghi chú:

$$p_{ij}^{(1)} = p_{ij}; \quad \forall i, j \in \mathbb{Z}.$$

$$p_{ij}^{(0)} = \delta_{ij} = \begin{cases} 1, & \text{nếu } i = j, \\ 0, & \text{nếu } i \neq j, \end{cases}$$

Trong đó δ_{ij} là ký hiệu **Kronecker.**

Phương trình **Chapman-Kolmogorov**

$$p_{ij}^{(m)} = \sum_{k \in \mathbb{Z}} p_{ik}^{(r)} p_{kj}^{(m-r)}$$

Từ đó ma trận xác suất chuyển m-bước của xích Markov là:

$$\mathbf{P}^{(m)} = \left(p_{ij}^{(m)}\right); \quad m = 0, 1, \dots$$

Vậy phương trình Chapman-Kolmogorov có thể viết dưới dạng:

$$\mathbf{P}^{(m)} = \mathbf{P}^{(r)}\mathbf{P}^{(m-r)}; \quad r = 1, 2, \dots, m.$$

Hay viết dưới dạng tích ma trận là:

$$\mathbf{P}^{(m)} = \mathbf{P}^m.$$

Điều kiện cân bằng và trạng thái hấp thụ Phân bố dừng: Điều kiện cân bằng tồn tại nếu xác suất trạng thái không thay đổi sau một số lượng lớn thời kỳ.

$$\pi(\text{thời kỳ tiếp}) = \pi(\text{thời kỳ này})$$

Hay

$$\pi = \pi\mathbf{P} \quad (*)$$

π được gọi là phân phối dừng của xích Markov.

Nhận xét:

- Nếu thực hiện phép chuyển vị công thức (*) thì ta có:

$$\mathbf{P}^T {p_i}^T = {p_i}^T, \quad \text{nghĩa là } {p_i}^T \text{ là vector riêng của } \mathbf{P}^T \text{ ứng với trị riêng } 1.$$

- Phân bố dừng chỉ phụ thuộc ma trận xác suất chuyển, không phụ thuộc phân bố ban đầu.
- Nếu phân phối dừng π không có xác suất 0 thì chuỗi Markov được gọi là có tính **Ergodic** và π là ma trận phân bố Ergodic.

C. Monte Carlo

Thuật ngữ "Monte Carlo" (tên một sòng bài ở Monaco) được sử dụng lần đầu bởi Metropolis (Los Alamos, 1947). Mô phỏng Monte Carlo là một kỹ thuật toán học dự đoán kết quả có thể xảy ra của một sự kiện không chắc chắn.

Trước đó, phương pháp ngẫu nhiên đã được sử dụng để tính số Pi (1901), hoặc tính tích phân, bằng cách lấy mẫu theo phân bố đều.

Kỹ thuật Monte Carlo bao gồm ba bước cơ bản:

- Thiết lập mô hình dự đoán: xác định biến phụ thuộc cần dự đoán và các biến độc lập/thuộc tính.
- Xác định phân phối xác suất của các biến độc lập: Sử dụng dữ liệu lịch sử và/hoặc đánh giá chủ quan của nhà phân tích để xác định một loạt các giá trị có thể xảy ra và chỉ định trọng số xác suất cho từng giá trị.
- Chạy mô phỏng lặp đi lặp lại, tạo ra các giá trị ngẫu nhiên của các biến độc lập. Làm điều này cho đến khi thu thập đủ kết quả để tạo thành một mẫu đại diện cho số lượng gần như vô hạn các kết hợp có thể có.

Phương pháp mô phỏng Monte Carlo là một phương pháp mô phỏng bằng xác suất. PP chủ yếu dựa trên hai luật quan trong của xác suất là luật số lớn và luật số lớn yếu.

Cho dãy $\{X_n\}$ các biến ngẫu nhiên độc lập có cùng phân phối với $E(X_i^2) < \infty$, khi đó ta có

$$\overline{X}_n \equiv \frac{X_1 + X_2 + \dots + X_n}{n} \xrightarrow{P} EX_1.$$

Luật yếu số lớn dạng Bernoulli

Nếu ta tiến hành n phép thử độc lập, với p là xác suất xuất hiện biến cố A trong mỗi phép thử và $\overline{X}_n$ là tần số xuất hiện biến cố đó trong n phép thử. Khi đó ta sẽ có

$$\overline{X}_n \xrightarrow{P} p.$$

IV. Phương Pháp Thực Hiện

A. Naive Bayes

Áp dụng **Định lý Bayes:**

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

Trong đó:

- $P(Y|X)$: Xác suất hậu nghiệm (posterior) của lớp Y (Rain/No rain) khi biết các đặc trưng X
- $P(X|Y)$: Xác suất của các đặc trưng X khi thuộc lớp Y
- $P(Y)$: Xác suất tiên nghiệm (prior) của lớp Y
- $P(X)$: Xác suất tiên nghiệm (prior) của đặc trưng X

Giả định **độc lập và xác suất có điều kiện**

Với giả định rằng các đặc trưng trong dữ liệu là độc lập, xác suất có điều kiện của các đặc trưng của trong lớp Y có thể được biểu diễn dưới dạng tích của từng xác suất thành phần:

$$P(X \mid Y) = P(X_1 \mid Y) \cdot P(X_2 \mid Y) \dots P(X_n \mid Y)$$

Với $X_1, X_2, \dots, X_n$ là các đặc trưng.

Hàm mật độ xác suất Gaussian

Trong bài toán này, các đặc trưng như nhiệt độ, độ ẩm, tốc độ gió là biến liên tục. Để tính xác suất của các đặc trưng liên tục, mô hình sử dụng phân phối Gaussian

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Xác suất hậu nghiệm

* Dựa vào giả định độc lập của Naïve Bayes, ta tính xác suất hậu nghiệm:

$$P(Y \mid x) = P(Y) \cdot \prod_{j=1}^{n} P(x_j \mid Y)$$

* Trong đó $P(x_j \mid Y)$ là xác suất có điều kiện của từng đặc điểm x_j khi thuộc lớp Y, và n là tổng số đặc điểm trong tập dữ liệu.

* Tuy nhiên, để tránh các lỗi về số vô cùng khi nhân các xác suất nhỏ, ta sử dụng log của công thức trên. Công thức xác suất hậu nghiệm sau khi áp dụng logarit sẽ là:

$$\log(P(Y \mid x)) = \log(P(Y)) + \sum_{j=1}^{n} \log\left(P(x_j \mid Y)\right)$$

Quyết định phân loại Để xác định lớp của một quan sát mới, mô hình chọn lớp Y có xác suất hậu nghiệm cao nhất:

$$Y = \arg\max P(Y \mid X) = \arg\max \left(P(Y) \cdot \prod_{i=1}^{n} P(x_i \mid Y) \right)$$

Quá trình thực hiện bài toán

Bước 1: Đọc và tiền xử lý dữ liệu bằng cách mã hóa các đặc trưng không ở dạng số học

Bước 2: Phân chia dữ liệu thành tập huấn luyện và tập test với tỉ lệ là 70:30. Cột 'RainTomorrow' được chọn làm biến phụ thuộc Y và các cột còn lại làm biến độc lập X

Bước 3: Xây dựng Classifier Naïve Bayes

- Tạo class với các thuộc tính cho xác suất tiên nghiệm (priors), giá trị trung bình (means) và phương sai (variances) của từng đặc trưng trong mỗi lớp
- Trong hàm fit: Tính xác suất tiên nghiệm, trung bình và phương sai của từng đặc điểm trong mỗi lớp, có thêm giá trị được gọi là "trơn" (smoothing) để tránh trường hợp chia cho 0

Bước 4: Tính xác suất hậu nghiệm và dự đoán

- Hàm predict_proba: Tính log xác suất hậu nghiệm cho mỗi lớp với từng mẫu trong tập test
- Hàm predict: Dựa trên xác suất hậu nghiệm đã tính, chọn lớp có xác suất cao nhất làm dự đoán

Bước 5: Đánh giá mô hình bằng cách dự đoán tập test bằng hàm predict và tính accuracy và confusion matrix để đánh giá mô hình

Bước 6: Kiểm thử trên dữ liệu thời tiết mới với các đặc trưng được mã hóa tương tự như trong quá trình huấn luyện và xuất ra dự đoán bằng phương pháp softmax.

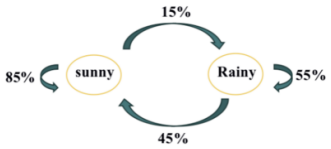
Công thức softmax được sử dụng để chuyển đổi các xác suất thô mà mô hình Naïve Bayes tính được thành xác suất chuẩn hóa để so sánh cho các lớp No Rain, Rain:

$$\text{Softmax}(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

B. Markov Chains

Áp dụng **Ma Trận Chuyển Tiếp** và **Dự Báo Trạng Thái:** Ma trận chuyển tiếp chứa xác suất chuyển từ trạng thái hiện tại sang trạng thái tiếp theo.

Ví dụ: Nếu thời tiết hôm nay là nắng, có 85% khả năng ngày mai cũng sẽ là nắng và 15% là mưa. Ma trận chuyển tiếp cho ví dụ này có dạng:



Với trạng thái $S = [\text{Nắng}, \text{Mưa}]$. Khi đó, trạng thái dự đoán sau một ngày là $S_1 = S_0 \cdot P$ và tương tự cho các ngày tiếp theo.

Trạng Thái Ổn Định:

Trong dài hạn, chuỗi Markov sẽ hội tụ đến một trạng thái ổn định không phụ thuộc vào trạng thái ban đầu.

Cách tính dự đoán sau N ngày:

Ta có thể tính bằng việc ta lấy dữ liệu cuối cùng được ghi nhận. Trong trường hợp bên dưới là ngày cuối cùng có ghi nhận mưa hay không.

VD: Dựa theo tỉ lệ trên, xác suất 3 ngày sau nếu ngày cuối là nắng:

$$S_3 = S_0 \cdot (P)^3$$

$$= \begin{bmatrix} 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0.85 & 0.15 \\ 0.45 & 0.55 \end{bmatrix}^3 = \begin{bmatrix} 0.766 & 0.236 \end{bmatrix}$$

Quá trình thực hiện bài toán:

Bước 1: Chuẩn Bị và Tiền Xử Lý Dữ Liệu

- Thu thập và tiền xử lý dữ liệu thời tiết, mã hóa các biến và tạo chuỗi trạng thái như "Có mưa"và "Không mưa".

Bước 2: Xây Dựng Ma Trận Chuyển Tiếp

- Sử dụng dữ liệu lịch sử để xác định xác suất chuyển tiếp giữa các trạng thái. Mỗi phần tử trong ma trận là xác suất chuyển tiếp từ trạng thái này sang trạng thái khác.

Bước 3: Huấn Luyện Mô Hình Markov

- Sử dụng ma trận chuyển tiếp để tính toán trạng thái ổn định. Điều này cho phép chúng ta ước tính xác suất dài hạn của các trạng thái thời tiết.

Bước 4: Dự Báo Kết Quả

- Sử dụng trạng thái ban đầu và ma trận chuyển tiếp để dự đoán thời tiết trong các khoảng thời gian tương lai.

C. Monte Carlo

Bài toán dự đoán mưa với biến mục tiêu

Bài toán đặt ra là dự đoán khả năng có mưa vào ngày mai, biểu thị qua biến mục tiêu "RainTomorrow", với hai giá trị "Có (1)" và "Không (0)".

Bài toán thuộc dạng phân loại nhị phân, với các đặc trưng đầu vào là các biến thời tiết như nhiệt độ, độ ẩm, áp suất, tốc độ gió, ...

Phương pháp Hồi quy Logistic

Hồi quy Logistic là một phương pháp học máy có giám sát phổ biến cho các bài toán phân loại nhị phân. Phương pháp này mô hình hóa xác suất xảy ra một sự kiện (trong trường hợp này là mưa) dựa trên một hàm logistic:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

Trong đó:
- $P(y = 1|X)$ là xác suất xảy ra mưa (RainTomorrow = 1).
- X là các biến đặc trưng.
- β là các hệ số hồi quy cần tìm.

Mô phỏng Monte Carlo

Mô phỏng Monte Carlo là một kỹ thuật mô phỏng ngẫu nhiên dựa trên phân phối xác suất của các biến đầu vào để tạo ra các kịch bản ngẫu nhiên khác nhau, giúp đánh giá độ ổn định và độ tin cậy của mô hình.

Trong phạm vi bài toán này, mô phỏng Monte Carlo được áp dụng bằng cách:

- Sinh các mẫu ngẫu nhiên từ phân phối của từng đặc trưng, dựa trên trung bình (mean) và độ lệch chuẩn (std) của chúng theo từng giá trị của biến RainTomorrow.
- Lặp lại nhiều lần quá trình huấn luyện và đánh giá mô hình để thu thập phân phối của độ chính xác, từ đó đánh giá độ ổn định của mô hình dự đoán.

Quá trình thực hiện bài toán

Bước 1: Tạo và chia sẻ dữ liệu

- Đọc và tiền xử lý dữ liệu bằng cách mã hóa các đặc trưng không ở dạng số học.
- Tập dữ liệu gốc được chia thành tập huấn luyện và kiểm thử với tỷ lệ 8:2 để chuẩn bị cho mô hình hồi quy logistic.

Bước 2: Hàm `monte_carlo_simulation`

Hàm `monte_carlo_simulation` chịu trách nhiệm sinh các tập dữ liệu mô phỏng để huấn luyện mô hình:

- **Phân tích dữ liệu ban đầu**: tính toán trung bình và độ lệch chuẩn của mỗi đặc trưng, chia thành hai nhóm theo biến RainTomorrow = 1 và RainTomorrow = 0.
- **Sinh mẫu ngẫu nhiên có điều kiện**:
  - Sử dụng phân phối chuẩn để sinh các giá trị mới cho từng đặc trưng, theo trung bình và độ lệch chuẩn đã tính ở trên.
  - Giá trị RainTomorrow cho mỗi hàng được sinh ngẫu nhiên theo xác suất xuất hiện của nó trong dữ liệu gốc.
- **Kết quả**: Hàm trả về danh sách các tập dữ liệu mô phỏng, mỗi tập mang tính ngẫu nhiên và tuân theo phân phối thống kê của dữ liệu gốc.

Bước 3: Vòng lặp Monte Carlo

Chạy nhiều vòng lặp (100 lần) để tạo các kịch bản khác nhau mô phỏng độc lập.

Trong mỗi vòng lặp:

- Gọi hàm `monte_carlo_simulation` để tạo 100 tập dữ liệu mô phỏng từ tập huấn luyện gốc.
- Kết hợp dữ liệu mô phỏng thành một tập huấn luyện đầy đủ, với các đặc trưng và biến mục tiêu RainTomorrow.
- Huấn luyện mô hình hồi quy logistic: Mô hình được huấn luyện trên tập dữ liệu mô phỏng để tìm ra các hệ số phù hợp với dữ liệu ngẫu nhiên này.
- Đánh giá mô hình trên tập kiểm thử: Sau khi huấn luyện, mô hình sẽ dự đoán trên tập kiểm thử gốc và tính toán các thông số đánh giá của dự đoán.

Bước 4: Phân tích kết quả mô phỏng Monte Carlo

Tính **trung bình** và **độ lệch chuẩn** của các thông số đánh giá:

Sau khi hoàn thành tất cả vòng lặp, chương trình tính trung bình và độ lệch chuẩn của các thông số đánh giá qua các kịch bản khác nhau.

V. Kết Quả và Kết Luận

A. Naive Bayes

Kết quả đánh giá mô hình Naive Bayes trong việc dự đoán khả năng mưa cho thấy một số thông số quan trọng như sau:

- **Độ Chính Xác (Accuracy)**: Mô hình đạt được độ chính xác 81.26%, cho thấy một tỷ lệ cao trong việc dự đoán đúng các trường hợp có và không có mưa. Điều này chỉ ra rằng mô hình hoạt động tốt trong việc phân loại các trường hợp.
- **Ma Trận Nhầm Lẫn (Confusion matrix)**:
  - *True Positives (TP)*: 5,423 trường hợp đã được dự đoán đúng là có mưa.
  - *False Positives (FP)*: 3,920 trường hợp bị dự đoán sai là có mưa.
  - *False Negatives (FN)*: 3,994 trường hợp bị dự đoán sai là không có mưa.
  - *True Negatives (TN)*: 28,900 trường hợp đã được dự đoán đúng là không có mưa.

  Ma trận nhầm lẫn cho thấy rằng mô hình có một số nhầm lẫn giữa các trường hợp có mưa và không có mưa, với số lượng dự đoán sai tương đối cao.
- **Độ Nhạy (Recall)**: Mô hình đạt được độ nhạy 57.59%, cho thấy khả năng phát hiện các trường hợp thực sự có mưa là tương đối hạn chế. Điều này chỉ ra rằng mô hình có thể bỏ lỡ một số trường hợp mưa.
- **Độ Chính Xác (Precision)**: Độ chính xác của mô hình là 58.04%, cho thấy rằng trong số những trường hợp mà mô hình dự đoán có mưa, chỉ khoảng hơn một nửa là chính xác. Điều này đồng nghĩa với việc mô hình có thể tạo ra nhiều dự đoán sai rằng có mưa.

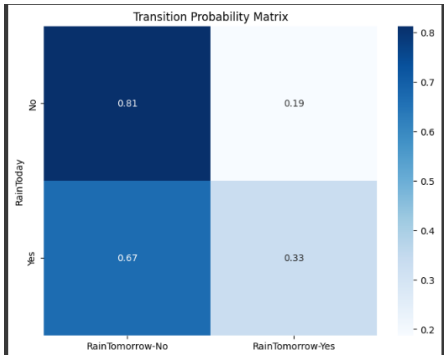
Lý do: Độ nhạy và độ chính xác có kết quả khá thấp, nguyên nhân là vì ở bộ dữ liệu, phần lớn kết quả là No Rain, chỉ có một phần nhỏ là Rain dẫn tới việc hai độ đo trên cho ra kết quả không tốt.

Tóm lại, mô hình Naive Bayes đã thể hiện được hiệu suất đáng khích lệ trong việc dự đoán khả năng mưa, nhưng vẫn còn một số điểm cần cải thiện để đạt được độ chính xác và độ nhạy cao hơn trong các dự đoán thực tế.

B. Markov Chains

Sau khi thực hiện với 100000 lần lặp dựa trên xác suất trong ma trận chuyển tiếp 'Matrix_T'. Mô hình in ra xác suất dự đoán tình hình thời tiết ngày mai dựa vào thông tin thời tiết ngày hôm nay.

Ma trận chuyển tiếp 'Matrix_T' được xác định như sau:



Trong đó:

- Khả năng ngày mai không mưa nếu hôm nay không mưa: 81.26%
- Khả năng ngày mai có mưa nếu hôm nay không mưa: 18.74%
- Khả năng ngày mai không mưa nếu hôm nay có mưa: 66.75%
- Khả năng ngày mai có mưa nếu hôm nay có mưa: 33.25%

Sau khi thực hiện mô phỏng và tính toán các chỉ số đánh giá, thu được kết quả như sau:

Confusion Matrix (Ma Trận Nhầm Lẫn):

- **True Negative (TN):** 6251 (Số lần mô hình dự đoán đúng rằng ngày mai không mưa)
- **False Positive (FP):** 1501 (Số lần mô hình dự đoán có mưa nhưng thực tế không mưa)
- **False Negative (FN):** 1501 (Số lần mô hình dự đoán không mưa nhưng thực tế có mưa)
- **True Positive (TP):** 747 (Số lần mô hình dự đoán đúng rằng ngày mai có mưa)

Kết quả thu được từ mô hình:

- **Accuracy:** 71.16%
- **Precision/Recall:** 34.16%

Tóm lại, Markov Chains cho thấy độ chính xác tổng thể của mô hình ở mức khá (71.16%), nhưng khả năng dự đoán ngày mai có mưa (Recall / Precision) còn hạn chế, chỉ đạt %. Điều này là do ma trận chuyển tiếp có xác suất chuyển tiếp từ "Có mưa"sang "Không mưa"và từ "Không mưa"sang "Có mưa"đều thấp, gây khó khăn cho mô hình trong việc phát hiện các ngày có mưa.

C. Monte Carlo

Kết quả đánh giá của mô hình trong việc dự khả năng mưa cho thấy các thông số như sau:

- **Độ chính xác tổng thể (Accuracy)**: Mô hình đạt được độ chính xác tổng thể 79%, cho thấy một tỷ lệ khá cao trong việc dự đoán đúng các trường hợp ngày có mưa và không mưa. Điều này chỉ ra rằng mô hình hoạt động khá tốt trong việc phân loại dữ liệu.

Lớp 1 (Có mưa):

- **Độ chính xác (Precision)**: mô hình đạt 53%, nghĩa là một tỷ lệ tương đối cao các cảnh báo sai (False Positives) – dự đoán có mưa nhưng thực tế không mưa.
- **Độ nhạy (Recall)**: mô hình đạt 64%, cho thấy khả năng phát hiện đúng ngày có mưa chưa cao.

Lớp 0 (Không có mưa):

- **Độ chính xác (Precision)**: mô hình đạt 89%, cho thấy trong số các dự đoán rằng sẽ không có mưa, số lượng cảnh báo sai là thấp. Mô hình dự đoán khá tốt cho các ngày không mưa.
- **Độ nhạy (Recall)**: mô hình đạt 84%, cho thấy khả năng phát hiện ngày thực tế không mưa là khá cao, nhưng vẫn còn một số ít trường hợp bị nhầm lẫn.

Với các thông số trên, có thể thấy ở việc dự đoán ngày có mưa của mô hình còn nhiều sai sót, một phần lớn đến từ lý do bộ dữ liệu ban đầu bị lệch về tập nhãn 0 (Không có mưa) dẫn đến việc kết quả không được đồng đều.

Tóm lại, phương pháp kết hợp giữa Mô phỏng Monte Carlo cùng với thuật toán Hồi quy Logistic đã thể hiện được hiệu suất đáng khích lệ trong việc dự đoán khả năng mưa, nhưng vẫn còn một số điểm cần cải thiện để đạt được độ chính xác và độ nhạy cao hơn trong các dự đoán thực tế.

VI. Phân Công Công Việc

Họ và Tên	Công việc	Mức độ
Phạm Đông Hưng	Thuyết trình, Report	100%
Lương Anh Huy	Thuyết trình, Slide, Nội dung	100%
Phan Công Minh	Thuyết trình, Slide, Nội dung	100%
Hồng Khải Nguyên	Thuyết trình, Nội dung	100%

References

Tài liệu

[1] https://tuanio.github.io/posts/markov-chain-va-bai-toan-sang-nay-an-gi/
[2] https://medium.com/@rangavamsi5/na%C3%AFve-bayes-algorithm-implementation-from-scratch-in-python-7b2cc39268b9
[3] https://machinelearningcoban.com/2017/08/08/nbc/
[4] https://www.geeksforgeeks.org/what-is-monte-carlo-simulation/
[5] chatGPT