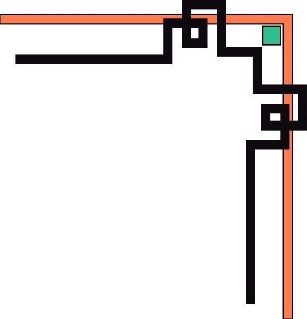
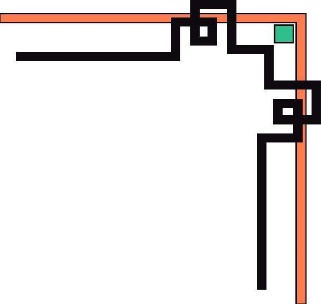
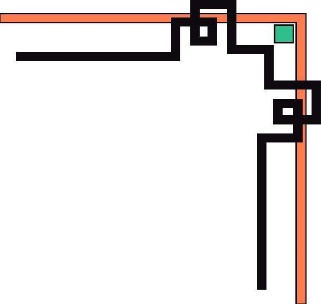
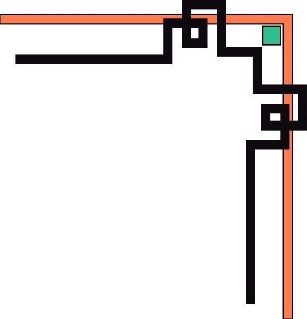
**BỘ GIÁO DỤC VÀ ĐÀO TẠO**



**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TPHCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**MÔN: KHAI PHÁ DỮ LIỆU**

**BÁO CÁO**

**ĐỀ TÀI:**

**DỰ BÁO THỜI TIẾT DỰA TRÊN QUAN SÁT HẰNG NGÀY TỪ CÁC TRẠM THỜI TIẾT Ở ÚC**

**GVHD: Ths. QUÁCH ĐÌNH HOÀNG**

**SVTH: Trần Thành Quang 19133047**

**Hoàng Minh Nhật 19133042**

**Bùi Quốc Kiệt 19133030**

**Đỗ Minh Đức 19133021**

**TP. Hồ Chí Minh, ngày 11 tháng 6 năm 2022**

**MỤC LỤC**

[**1. TÓM TẮT (ABSTRACT)** 3](#_Toc105942534)

[**2. GIỚI THIỆU (INTRODUCTION)** 4](#_Toc105942535)

[**3. DỮ LIỆU (DATA)** 5](#_Toc105942536)

[**4. TIỀN XỬ LÝ DỮ LIỆU** 9](#_Toc105942537)

[**5. MÔ HÌNH HÓA DỮ LIỆU (DATA MODELING)** 12](#_Toc105942538)

[**5.1 Lựa chọn độ đo và biến quan trọng** 12](#_Toc105942539)

[**5.1.1 . Lựa chọn biến quan trọng** 12](#_Toc105942540)

[**5.1.2 . Lựa chọn độ đo** 12](#_Toc105942541)

[**5.2. Decision tree** 13](#_Toc105942542)

[**5.3 Support vector Machine (SVM)** 15](#_Toc105942543)

[**5.4 Logistic Regression** 16](#_Toc105942544)

[**5.5 Random forest** 16](#_Toc105942545)

[**6. THỰC NGHIỆM, KẾT QUẢ, VÀ THẢO LUẬN** 18](#_Toc105942546)

[**6.1. Dự đoán sử dụng thuật toán Decission tree** 18](#_Toc105942547)

[**6.2 Dự đoán sử dụng thuật toán SVM** 18](#_Toc105942548)

[**6.3. Dự đoán sử dụng thuật toán Logistic regression** 20](#_Toc105942549)

[**6.4. Dự đoán sử dụng thuật toán Random-forest.** 20](#_Toc105942550)

[**7. KẾT LUẬN (CONCLISION)** 22](#_Toc105942551)

[**8. PHỤ LỤC (APPENDICES)** 23](#_Toc105942552)

[**9. ĐÓNG GÓP (CONTRIBUTIONS)** 24](#_Toc105942553)

[**TÀI LIỆU THAM KHẢO** 25](#_Toc105942554)

# **1. TÓM TẮT (ABSTRACT)**

Quan sát thời tiết hàng ngày từ nhiều địa điểm trên khắp nước Úc, được lấy từ Cục Khí tượng Liên bang Úc và được xử lý để tạo ra bộ dữ liệu mẫu lớn này để minh họa phân tích, khai thác dữ liệu và khoa học dữ liệu bằng R và Rattle.

Dữ liệu đã được xử lý để cung cấp một biến số **Raintomorrow** (có mưa vào ngày hôm sau hay không) và một biến số rủi ro rủi **risk\_mm** (lượng mưa là bao nhiêu được ghi lại tính bằng milimet). Các biến đổi khác nhau được thực hiện trên dữ liệu.

**Phương pháp thực hiện :**

* + Xử lý dữ liệu : làm sạch giá trị NA , NULL, String ….
  + Phân tích vẽ biểu đồ trực quan hóa dữ liệu
  + Thống kê tóm tắt dữ liệu
  + Xây dựng mô hình dữ đoán biến đầu ra :
    - Chia tập dữ liệu thành 2 phần : 80% train , 20 % test .
    - Sử dụng thuật toán (decision tree , randomforest , SVM , Logistic regresstion)
  + Đánh giá các mô hình dự đoán :
    - Confusion matrix
    - Precision , Recall , F1-Score

**Mục tiêu :**

Mục tiêu của bài toán là dự đoán được thời tiết ngày tiếp theo ở Úc (có mưa hay không ).

# **2. GIỚI THIỆU (INTRODUCTION)**

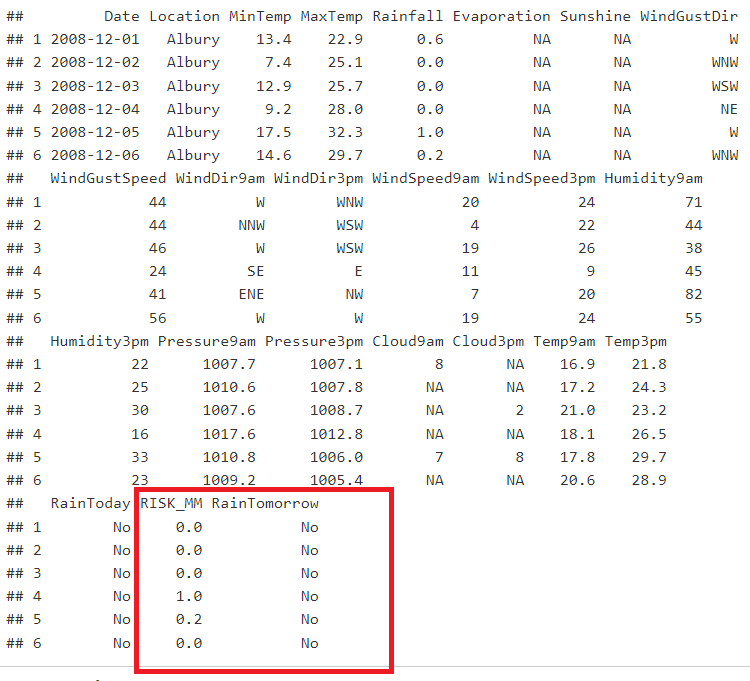
Dự báo thời tiết là một ngành ứng dụng của khoa học và công nghệ để tiên đoán trạng thái và vị trí của bầu khí quyển trong tương lai gần. Loài người đã nỗ lực dự báo thời tiết một cách không chính thức từ nhiều thiên niên kỳ trước, và việc dự báo thời tiết một cách chính thức bắt đầu từ thế kỷ mười chín.

Công tác dự báo thời tiết được thực hiện bằng cách thu thập số liệu về trạng thái hiện tại của bầu khí quyển và áp dụng những hiểu biết khoa học về các quá trình của khí quyển để tiên đoán sự tiến triển của khí quyển.

Dự báo thời tiết có một vai trò to lớn trong đời sống kinh tế, xã hội. Dự báo đúng sẽ giúp con người đưa ra những quyết định đúng đắn. Đề tài nghiên cứu và áp dụng các mô hình cho việc dự báo thời tiết như :

* Decision tree
* Random Forest
* Support Vector Machine
* Logistic Regression

**Output** của bài toán là biến kết quả **Raintomorrow & RISK\_MM** , **input** là các biến còn lại . Tuy nhiên trong đề tài này sẽ không sử dụng biến risk\_mm.



# **3. DỮ LIỆU (DATA)**

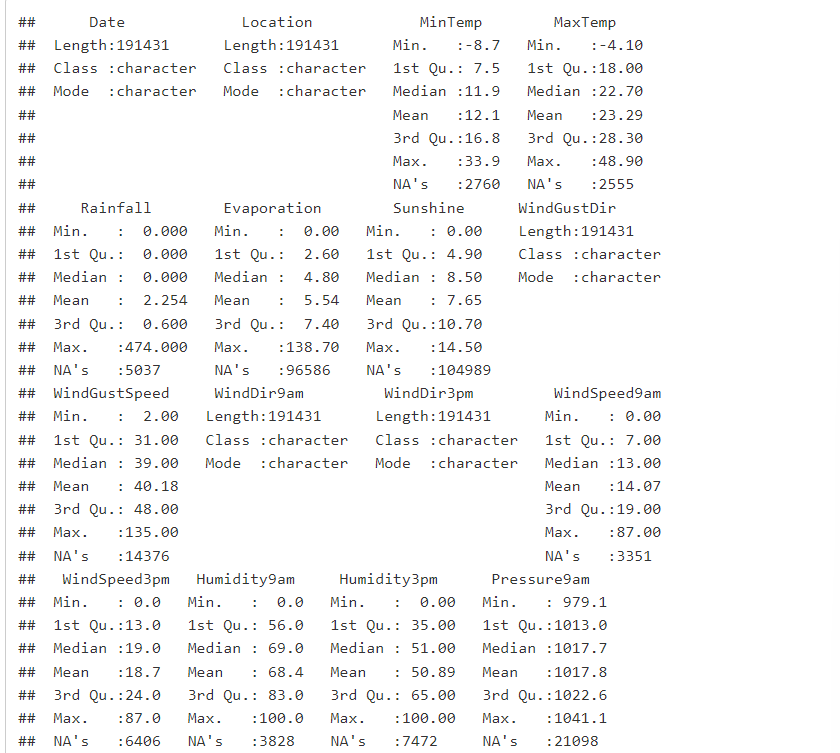
**Nguồn :** Bộ dữ liệu nguồn đến từ Cục Khí tượng Khối thịnh vượng chung Úc. Cục đã cho phép sử dụng dữ liệu với Cục Khí tượng được thừa nhận là nguồn của dữ liệu, theo email từ Cathy Toby (C.Toby@bom.gov.au) thuộc Dịch vụ Thông tin Khí hậu của Trung tâm CLimate Quốc gia, 17 Tháng 12 năm 2008.

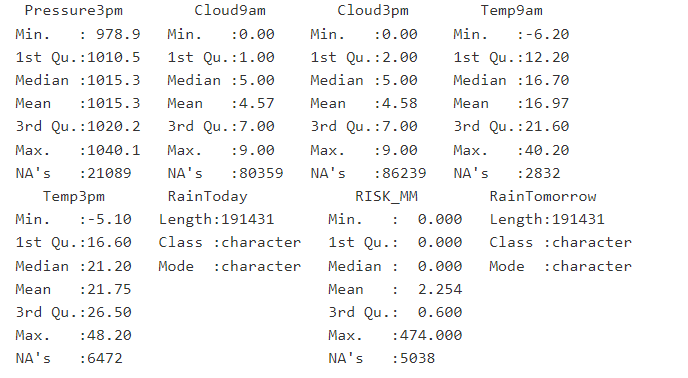
**Link nguồn :** <https://www.rdocumentation.org/packages/rattle/versions/5.4.0/topics/weatherAUS>

**Ý nghĩa các biến :**

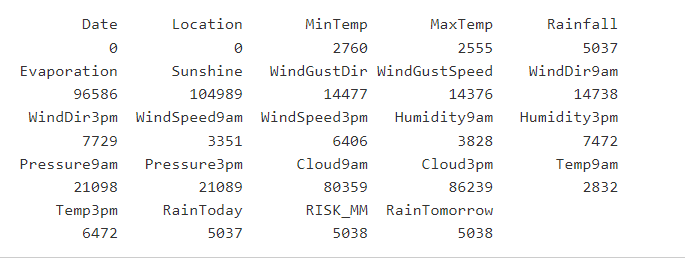
* **Date**: Ngày quan sát (một đối tượng Date).
* **Location**: Tên thông thường của vị trí của trạm thời tiết.
* **MinTemp**: Nhiệt độ tối thiểu tính bằng độ C.
* **MaxTemp**: Nhiệt độ tối đa tính bằng độ C.
* **Rainfall**: Lượng mưa được ghi lại trong ngày tính bằng mm.
* **Evaporation** : Cái gọi là độ bốc hơi của chảo loại A (mm) trong 24 giờ đến 9 giờ sáng.
* **Sunshine** : Số giờ nắng sáng trong ngày.
* **WindGustDir** : Hướng gió giật mạnh nhất trong 24 giờ đến nửa đêm.
* **WindGustSpeed** : Tốc độ (km / h) của gió giật mạnh nhất trong 24 giờ đến nửa đêm.
* **Temp9am** : Nhiệt độ (độ C) lúc 9 giờ sáng.
* **RelHumid9am** : Độ ẩm tương đối (phần trăm) lúc 9 giờ sáng.
* **Cloud9am** : Một phần bầu trời bị mây che khuất lúc 9 giờ sáng. Điều này được đo bằng “oktas”, là một đơn vị của eigth. Nó ghi lại bao nhiêu vùng trời bị mây che khuất. Số đo 0 cho biết bầu trời hoàn toàn trong khi số 8 cho biết trời hoàn toàn u ám.
* **WindSpeed9am** : Tốc độ gió (km / giờ) trung bình trong 10 phút trước 9 giờ sáng.
* **Pressure9am** : Áp suất khí quyển (hpa) giảm xuống mức trung bình của mực nước biển lúc 9 giờ sáng.
* **Temp3pm** : Nhiệt độ (độ C) lúc 3 giờ chiều.
* **RelHumid3pm** : Độ ẩm tương đối (phần trăm) lúc 3 giờ chiều.
* **Cloud3pm** : Một phần bầu trời bị mây che khuất (trong “oktas”: phần tám) lúc 3 giờ chiều. Xem Cload9am để biết mô tả về các giá trị.
* **WindSpeed3pm** : Tốc độ gió (km / giờ) trung bình trong 10 phút trước 3 giờ chiều.
* **Pressure3pm** : Áp suất khí quyển (hpa) giảm xuống mức trung bình của mực nước biển lúc 3 giờ chiều.
* **ChangeTemp** : Thay đổi nhiệt độ.
* **ChangeTempDir** : Hướng thay đổi của nhiệt độ.
* **ChangeTempMag** : Độ lớn của sự thay đổi của nhiệt độ.
* **ChangeWindDirect** : Hướng gió thay đổi.
* **MaxWindPeriod** : Thời kỳ gió cực đại.
* **RainToday** : Số nguyên: 1 nếu lượng mưa (mm) trong 24 giờ đến 9 giờ sáng vượt quá 1 mm, nếu không thì 0.
* **TempRange** : Sự khác biệt giữa nhiệt độ tối thiểu và tối đa (độ C) trong 24 giờ đến 9 giờ sáng.
* **PressureChange** : Thay đổi áp suất.
* **RISK\_MM** : Lượng mưa. Một loại thước đo “rủi ro”.
* **RainTomorrow** : Biến mục tiêu. Ngày mai trời có mưa không?

**Tóm tắt dữ liệu**

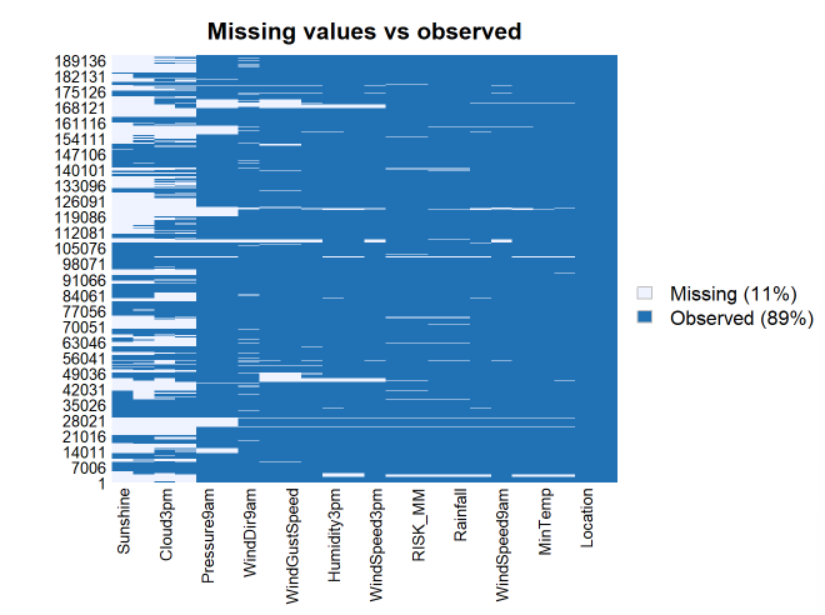




**Liệt kê tổng sô NA của từng biến**



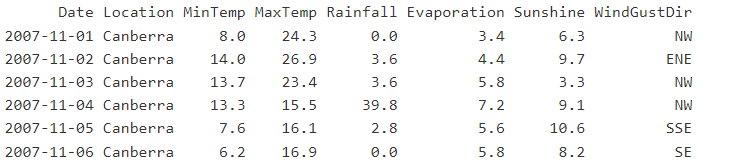
**Trực quan hóa các giá trị Missing**



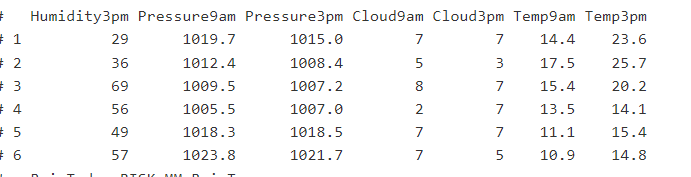
# **4. TIỀN XỬ LÝ DỮ LIỆU**

Loại bỏ các giá trị NA , Missing value

Sắp xếp tập dữ liệu tăng dần theo biến “Date” và hiển thị kết quả sau khi sắp xếp

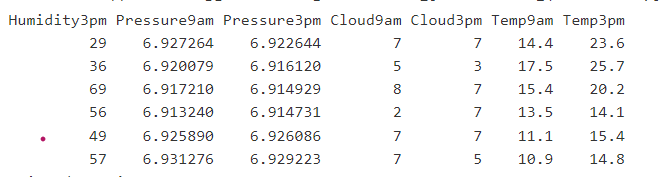


…

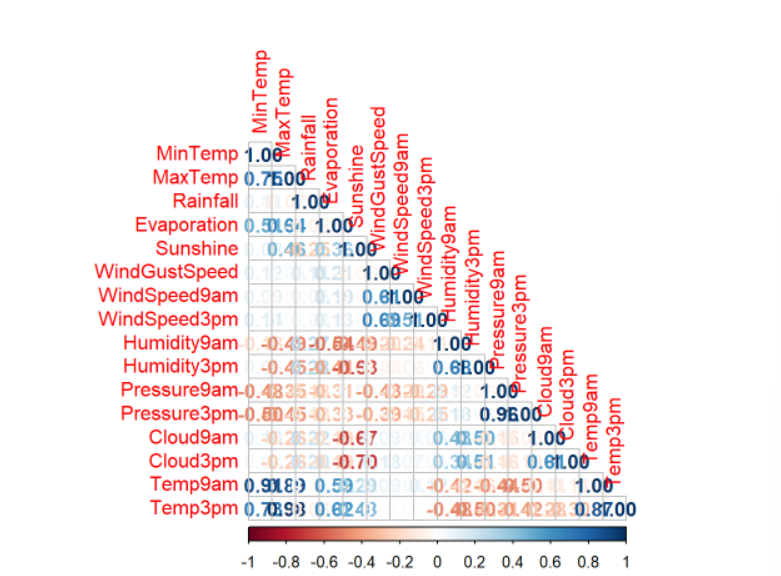


Như ở trên ta thấy có 2 cột là Pressure9am với Pressure3pm thì 2 cột này có miền giá trị hơi lớn so với các miền giá trị còn lại. Nếu ta không scale nó lại thì nó sẽ ảnh hưởng đến mô hình logistic regression nên ta sẽ lấy Log 2 biến này để giá trị nó nhỏ lại .Ngoài ra ta còn loại bỏ một số cột không được sử dụng trong model như Date , RISK\_MM

Kết quả sau khi log 2 biến Pressure9am với Pressure3pm

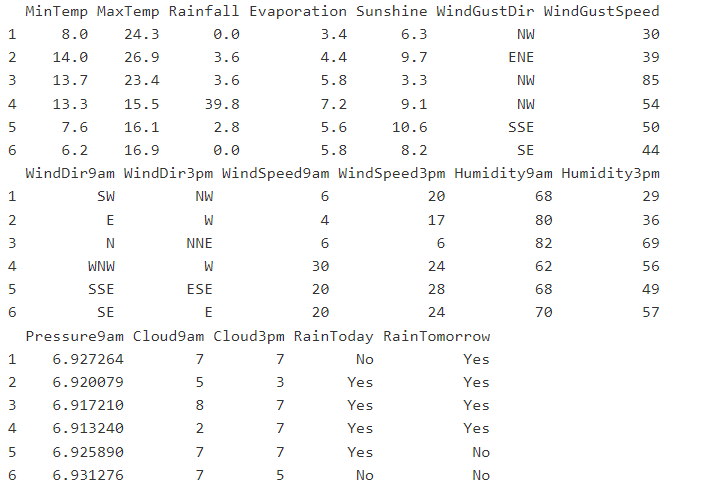


Tìm ra các biến có độ tương quan cao với nhau để loại bỏ để tránh ảnh hưởng đến model.

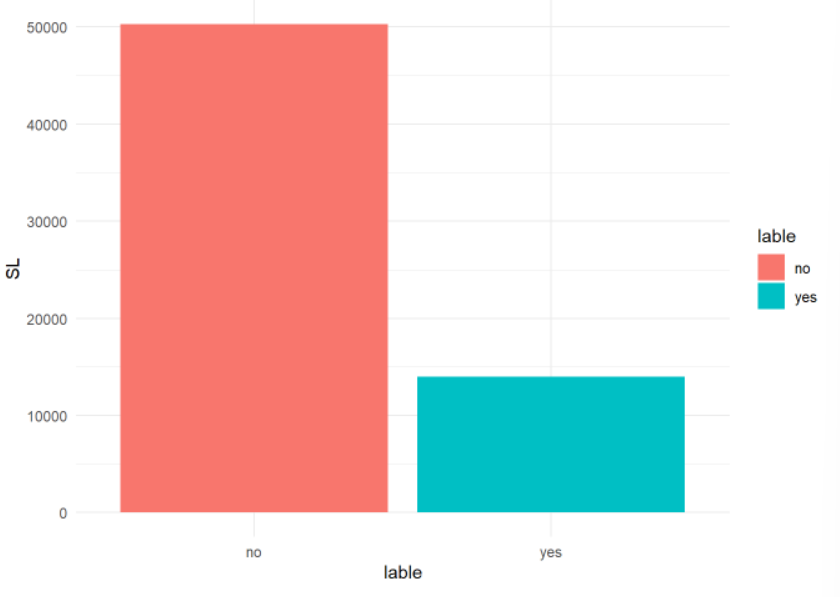


Nhận xét : ta thấy 3 biến là Temp9am , temp3pm, Pressure3pm lần lượt tương quan khá cao ( >0.9) với các biến MinTemp , Maxtemp , Pressur9am nên ta sẽ loại bỏ 3 biến này khỏi tập dữ liếu.

Sau khi làm sạch dữ liệu ta có có tập dữ liệu



Vẽ biểu đồ thể hiện sự chênh lệch giữa hai nhãn “YES” và “NO” của biến kết quả



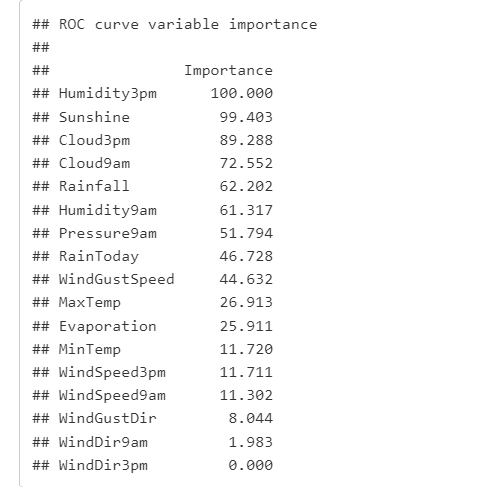
Nhận xét: ta thấy nhãn đầu ra của biến Raintomorrow đa số là nhãn “NO” chiếm hơn 3 lần so với nhãn “YES” nên ta kh sử dụng được độ đo Accuraty vì nếu sữ dụng độ đo này sẽ kh nhận xét được độ chính xác của mô hình. Ngoài ra việc âm tính giả cũng khá quan trọng nên nhóm quyết đinh chọn độ đo F-measure (F1) để đánh giá mô hình.

# **5. MÔ HÌNH HÓA DỮ LIỆU (DATA MODELING)**

## **5.1 Lựa chọn độ đo và biến quan trọng**

### **5.1.1 . Lựa chọn biến quan trọng**

Ở đay ta xây dựng một mô hình decision tree sau đó xem các biến quan trọng ảnh hưởng lớn đến model .

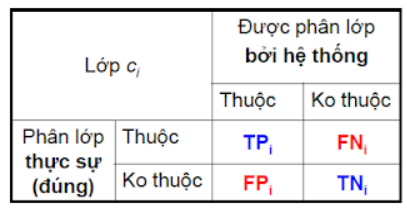


Nhận xét : ở đay ta có danh sách độ quan trọng của các biến ảnh hưởng đên model như trên . Nhóm sẽ lựa chọn 10 biến quan trọng nhất để xây dựng model dự đoán cho các thuật toán như Decision tree , SVM , logistic regession , random-forest.

### **5.1.2 . Lựa chọn độ đo**

Ở đây bài toán của nhóm là bài toán phân loại nên chúng ta sẽ tạo ra một Ma trận nhầm lẫn (Confusion matrix ) để sử dụng các độ đo Precision , recal , F-mease.

Ta có dạng của ma trận Confusion matrix.



Giải thích :

**TPi:** Số lượng các ví dụ thuộc lớp ci được phân loại chính xác vào lớp ci

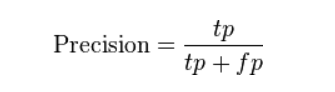
**FPi**: Số lượng các ví dụ không thuộc lớp ci bị phân loại nhầm vào lớp ci

**TNi**: Số lượng các ví dụ không thuộc lớp c**i** được phân loại (chính xác)

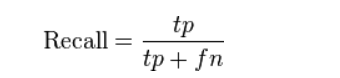
**FNi**: Số lượng các ví dụ thuộc lớp c**i** bị phân loại nhầm (vào các lớp khác c**i**)

Độ đo :

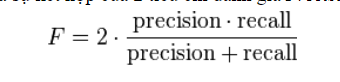
* Precision



* Recal



* F-mease : Tổng hợp của 2 tiêu chí Recal với Precision



## **5.2. Decision tree**

Cây quyết định ( Decision tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào một dãy các quy luật. Các thuộc tính của đối tượng có thể khác nhau như nhị phân (binary) , Định dang (nominal ) , thứ tự (ordinal) , số lượng (quantitative ) trong khi đó thuộc tính của lớp dự đoán phải có kiểu dữ liệu Binary hoặc ordinal.

Vd:ở đây ta ví dụ một ví dụ đơn giản , dựa vào thời tiết để xem ta có thế đi đá bóng hay không.

Thuộc tính đầu vào là :

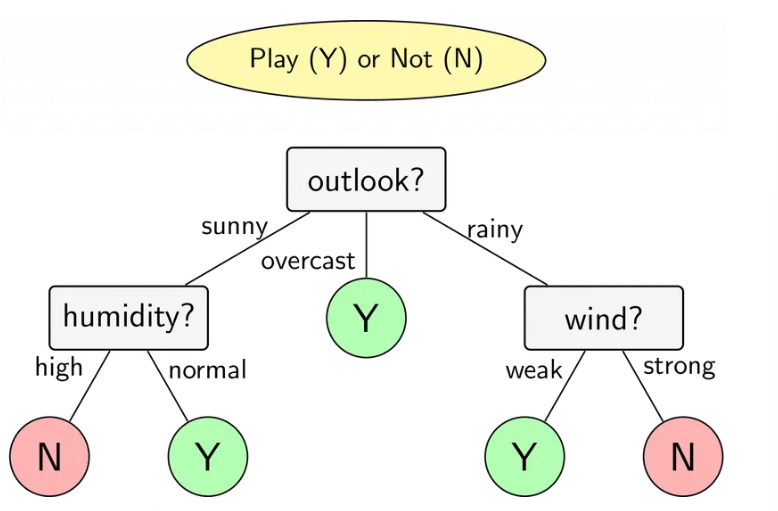
+ Thời tiết

+ Độ ẩm

+ Gió

Thuộc tính đầu ra : “YES” có đi đá bóng và “No” không đi đá bóng.

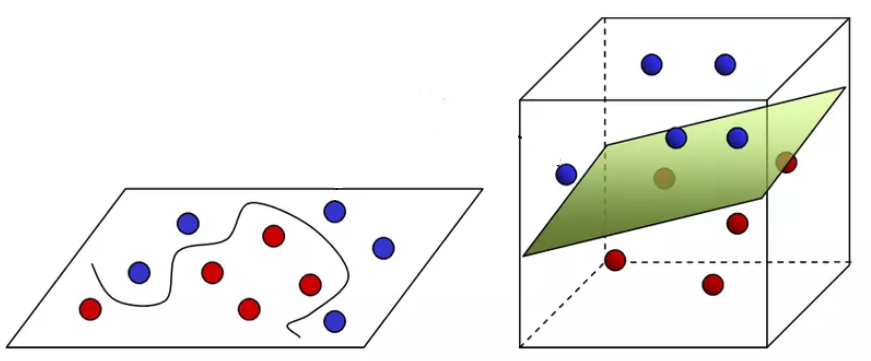
Ta sẽ lâp được cây như thế này.



Dựa vào mô hình đơn giản ở trên ta thấy nếu trời nắng, độ ẩm bình thường , khả năng em sẽ đi đá bóng sẽ cao.

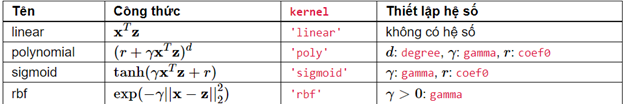
## **5.3 Support vector Machine (SVM)**

Support Vector Machine (SVM) là một thuật toán giám sát , nó có thể sử dụng cho cả việc hồi quy và phân loại. Mục tiêu của bài toán là tìm ra Hyper-plane (1) phân chia các lớp. Vấn đề đặt ra là có rất nhiều Hyper-plane như thế, vậy đâu mới là siêu phẳng phân chia dữ liệu tốt nhất. Margin (2) chính là câu trả lời, siêu phẳng có margin lớn nhất là thứ chúng ta cần tìm trong bài toán này.



Với một tập dữ liệu phân chia tốt, ta có thể sử dụng mô hình tuyến tính (Linear) để phân chia dữ liệu thành các lớp như hình bên phải. Nhưng với dữ liệu không có tính phân chia tốt như hình bên trái thì mô hình tuyến tính có vẻ không phân chia tốt. Vì thế chúng ta cần một bước biến đổi đại số, sử dụng mô hình phi tuyến tính (Non-Linear) để phân chia dữ liệu như sau :

Một số loại kernel :

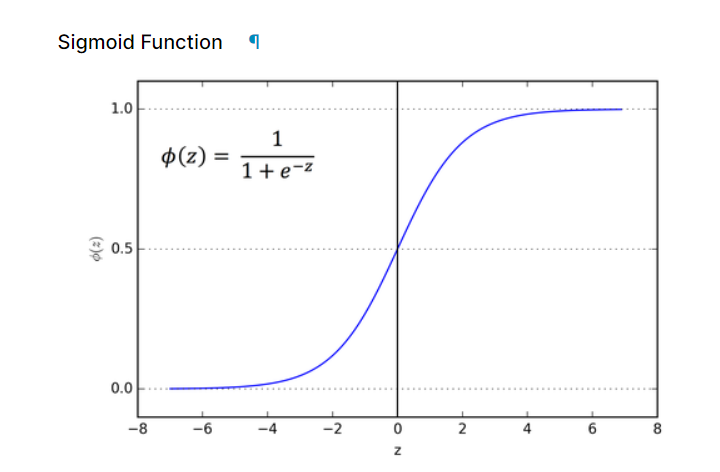


Với mỗi kernel sẽ có thông số riêng như trên hình. Mặc định coef0 bằng 0 , gamma bằng 1/dim , degree bằng 3.

## **5.4** **Logistic Regression**

Đây là một thuật toán phân loại học tập có giám sát được sử dụng để dự đoán các quan sát cho một tập hợp các lớp riêng biệt. Thực tế, nó được sử dụng để phân loại các quan sát thành các loại khác nhau. Do đó, đầu ra của nó là riêng biệt trong tự nhiên. Hồi quy logistic còn được gọi là hồi quy logit. Đây là một trong những thuật toán phân loại đơn giản, đơn giản và linh hoạt nhất được sử dụng để giải quyết các vấn đề phân loại.

Ta sử dụng hàm Sigmoid Fuction : hàm này sẽ dự đoán kết quả dự đoán có xác xuất từ 0-1 , sau khi ta dự đoán mô hình ta sẽ vẽ đồ thị Sigmoid function



Ở cái ranh giới quyết định ( decision boundary) thì nếu

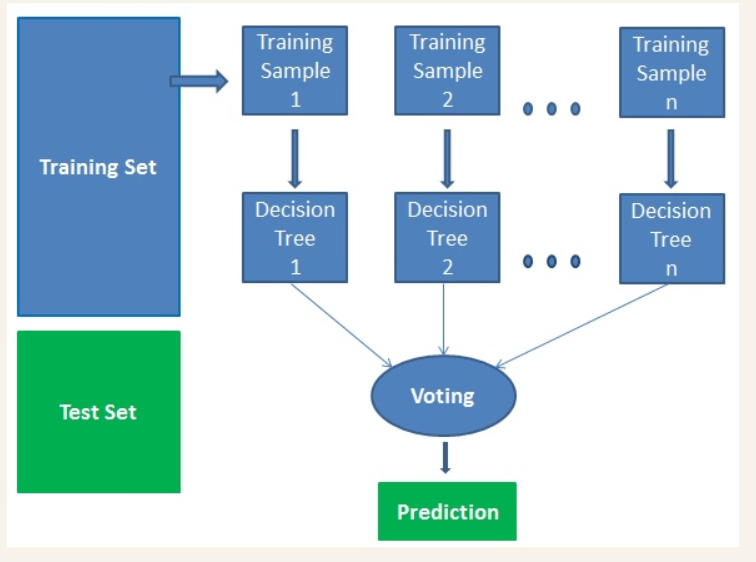
+ P>=0.5 ta sẽ gắn cho nó là “YES”

+ P <0.5 ta sẽ gắn cho nó là “No”

## **5.5 Random forest**

Random-forest là một thuật toán học có giám sát. Có thể sử dụng cho cả bài toán phân loại và hồi quy .

Thuật toán này sẽ tạo ra một khu rừng gồm rất nhiều câu quyết định dựa trên các mẫu ngẫu nhiên , để đưa ra một kết quả thì ta dựa vào cách bỏ phiểu giữa các cây trong khu rừng .



# **6. THỰC NGHIỆM, KẾT QUẢ, VÀ THẢO LUẬN**

## **6.1. Dự đoán sử dụng thuật toán Decission tree**

Đầu tiên ta chia tập dữ liệu thành hai phần là tập train và test với tỉ lệ: 80% tập train và 20% tập test. Ta sẽ lấy 80% từ cột 1 đến cột thứ 51383 làm train , và cột từ 51383 làm tập test . Vì ta lấy dữ liệu ở quá khứ để dữ đoán cho tương lại nên ta chia dữ liệu theo thứ tự như vậy.

Sau khi ta chia xong 2 tập train với test ta sẽ xây dựng model Decision tree với 10 biến quan trọng mà em đã trình bày ở trên . Sunshine , Humidity3pm , Cloud3pm , Cloud9am , Rainfall, Humidity9am ,Pressure9am , RainToday + WindGustSpeed + Evaporation

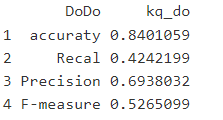
F-measure :

* Trên tập train : 0.5914329
* Trên tập test : 0.5265099

Nhận xét : ta thấy mô hình dự đoán có tỉ lệ chênh lệch giữa train và test cũng khá cao

Ngoài ra mô hình còn dự đoán khá thấy so với kì vọng mà nhóm đặt ra với mô hình decision tree.

Ngoài ra nhóm còn xuất ra thử kết quả của 4 độ đo trên test gồm Accruty , recal, Precsion

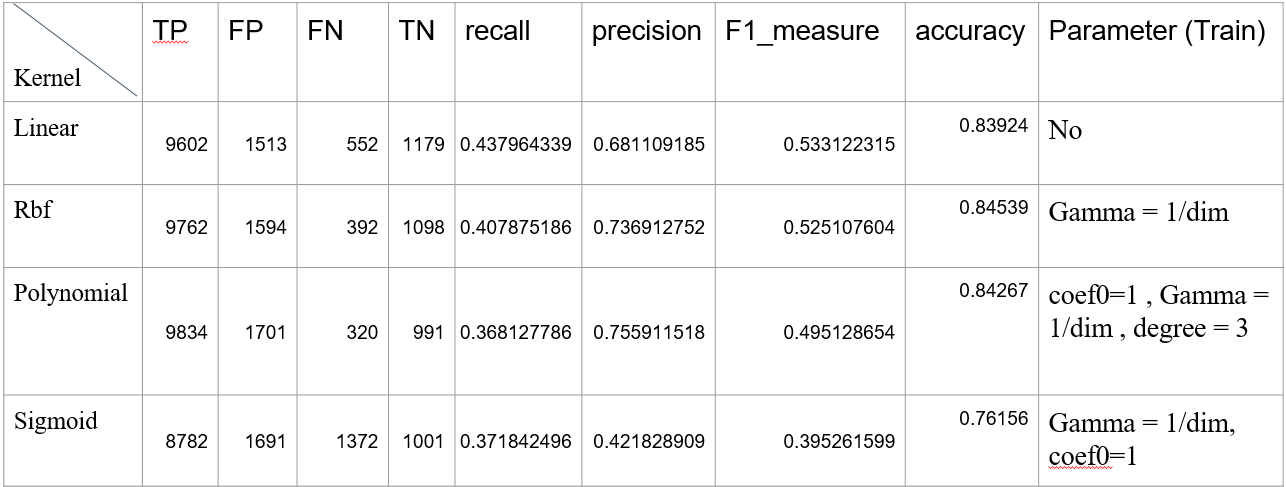


Nhận xét : ta thấy độ chính xác của nó khá cao 84% nhưng độ F1 của nó khá thấp.

## **6.2 Dự đoán sử dụng thuật toán SVM**

Sử dụng tập Train (80%) và Test (20%) như đã phân chia ở trên. Dùng thư viện e1071 cho thuật toán SVM. Train model với cost = 1 , SVM-Type = C-classification trên tập Train ( 80% ).

Kết quả dự đoán trên tập Test (20%) :



**Một số nhận xét với thuật toán SVM :**

Thuật toán svm có độ phức tạp khá cao , độ chính xác có thể được nâng lên khi tăng cost nhưng cũng đồng nghĩa với việc mất nhiều thời gian và tài nguyên tính toán hơn.

Với nhiều dạng dữ liệu khác nhau thì sẽ có kernel (hàm biến đổi đại số) phù hợp với loại dữ liệu đó hơn. có 4 dạng chính với thuật toán SVM trong thư viện e1071 : Linear , Polinomial , Sigmoid, Ratio Basis Function.

Thực nghiệm trên tập dữ liệu cho thấy kernel Ratio Basis Function cho độ chính xác cao nhất và F1-score thứ nhì. Kernel Linear có F1-Score (độ đo chính của nhóm) lớn nhất nên được chọn.

**Tối ưu :** Với kernel Linear không có tham số nên cách tối ưu là tăng cost tới mức phù hợp , không tăng quá cao vì sẽ bị overfitting trên tập train và kết quả không cao trên tập test. Khi tham số cost tăng lên thì model sẽ cố gắng fix các điểm nhiễu , vì thế Margin sẽ giảm. Để mô hình được Generalization hơn thì cost phải nhỏ , khi đó mô hình sẽ đạt được Margin lớn.

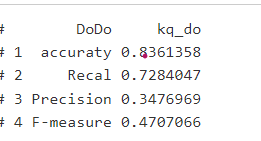
Tóm lại với cost nhỏ thì Margin lớn , độ khái quát cao. với cost tăng cao thì độ chính xác tăng lên , Margin giảm , chi phí tăng và tốc độ chậm hơn.

## **6.3. Dự đoán sử dụng thuật toán Logistic regression**

Kết quả F-measure

* trên tập train : 0.592
* trên tập test :0.471

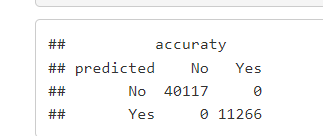
Nhận xét : kết qủa trên tập test quá thấp dưới 50% , nó chênh lệch với giữa tập train với tập test khá nhiều . Ngoài ra còn một số độ đo khác.

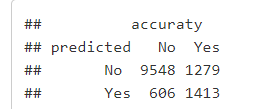


## **6.4. Dự đoán sử dụng thuật toán Random-forest.**

Trước tiên ta chia tập dữ liệu thành hai phần là tập train và test với tỉ lệ 8:2 train chiếm 80% và test chiếm 20%. Ta sẽ lấy 80% từ cột 1 đến cột thứ 51383 làm train , và cột từ 51383 làm tập test . Vì ta lấy dữ liệu ở ngày hôm nay để dự đoán cho tương lai có mưa hay không nên ta chia dữ liệu theo thứ tự như vậy.

Sau khi ta chia xong 2 tập train và test ta sẽ xây dựng model random-forest với các biến quan trọng mà chúng em đã đề cập ở trên như Sunshine , Humidity3pm , Cloud3pm , Cloud9am , Rainfall, Humidity9am ,Pressure9am , RainToday + WindGustSpeed + Evaporation

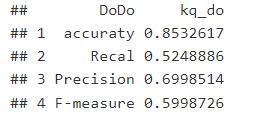
Trên tập train 

Trên tập test

Nhận xét : ta thấy mô hình dự đoán có tỉ lệ chênh lệch giữa train và test cũng không quá nhiều

Ngoài ra mô hình còn dự đoán khá thấp so với kỳ vọng mà nhóm đặt ra với mô hình random-forest.

Ngoài ra nhóm còn xuất ra thử kết quả của 4 độ đo trên test gồm Accuity , recal, Precision,F-measure



Độ chính xác của mô hình lên đến 85% nên khá cao mà ngược lại thì độ F1 có vẻ khá thấp so với mô hình.

# **7. KẾT LUẬN (CONCLISION)**

* SVM có độ phức tạp cao , tốc độ chậm , nhạy cảm với nhiễu.
* Decission tree tốc độ chậm , độ chính xác chưa cao.
* Logistic Regression tốc độ nhanh , độ chính xác chưa cao.
* Random Forest tốc độ nhanh , độ chính xác cao nhất nên là lựa chọn để phát triển .

**Hướng phát triển đề tài**

* Cải thiện phần tiền xử lý dữ liệu
* Tinh chỉnh các tham số cho các mô hình dự đoán được tốt hơn
* Tìm hiểu mô hình RNN(LSTM) xử lý dữ liệu chuỗi thời gian.

# **8. PHỤ LỤC (APPENDICES)**

1. Hyper-plane : Siêu phẳng có số chiều là (n-1) phân chia dữ liệu thành các lớp , trong đề tài này là hai lớp.
2. Margin : Khoảng cách giữa siêu phẳng và các quan sát gần siêu phẳng nhất (vectơ hỗ trợ).

# **9. ĐÓNG GÓP (CONTRIBUTIONS)**

|  |  |  |
| --- | --- | --- |
| **Thành viên** | **Nhiệm vụ** | **Mức độ hoàn thành** |
| Trần Thành Quang | Model Decision tree | 100% |
| Hoàng Minh Nhật | Model SVM | 100% |
| Đỗ Minh Đức | Model Logistic Regresstion | 100% |
| Bùi Quốc Kiệt | Model Random forest | 100% |

# **TÀI LIỆU THAM KHẢO**

* Quách Đình Hoàng, slide, video bài giảng khai phá dữ liệu, đại học Sư phạm Kỹ thuật thành phố Hồ Chí Minh, năm 2021.
* <https://machinelearningcoban.com/2016/12/28/linearregression/>
* <https://sites.google.com/site/diepnn80/datamininginfo/cacdodohaydungchobaitoanphanloai>

**Library:**

* library(Amelia)
* library(sqldf)
* library(tidyr)
* library(readr)
* library(dplyr)
* library(party)
* library(caret)
* library(ggplot2)
* library(e1071)
* library("randomForest")