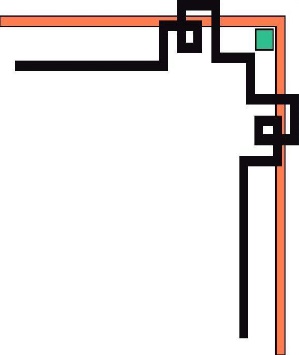
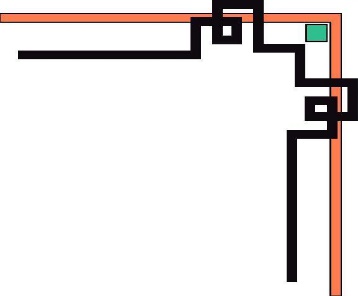
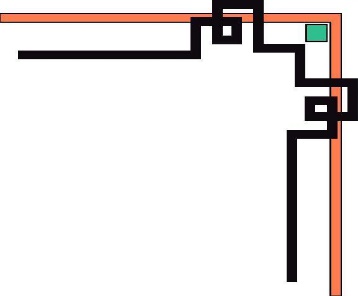
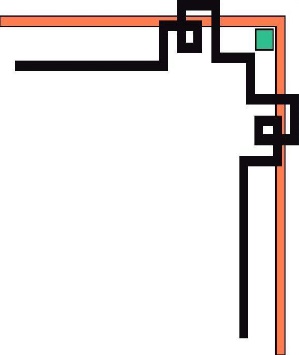
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**KHOA HỆ THỐNG THÔNG TIN**

-----🙞🙜🕮🙞🙜-----

**A picture containing text

Description automatically generated**

**Đề tài**

**Dự Báo Lượng Mưa Ở Australia**

**GIẢNG VIÊN HƯỚNG DẪN**

**Hà Lê Hoài Trung**

**NHÓM SINH VIÊN THỰC HIỆN**

Phan Cao Vũ– MSSV: 18521674

Contents

[CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI 3](#_Toc75898673)

[1.1 Lí do chọn đề tài 3](#_Toc75898674)

[1.2 Dữ Liệu 3](#_Toc75898675)

[1.2.1 Dataset Traffic Crashes – Crashes 3](#_Toc75898676)

[2.4 Xóa những thuộc tính không cần thiết : 6](#_Toc75898677)

[2.4.1 Xóa những thuộc tính trên file không cần thiết : 6](#_Toc75898678)

[Chương 3 : Tiền xữ lý dữ liệu 8](#_Toc75898679)

[3.1 xữ lí dữ liệu còn thiếu 8](#_Toc75898680)

[3.2 Kết quả kiểm tra các giá trị null trong bảng dữ liệu 10](#_Toc75898681)

[Chương 4 : Trực quan hóa dữ liệu 10](#_Toc75898682)

[Chương 5 : chạy thuật toán 18](#_Toc75898683)

[5.1. Kiểm tra dữ liệu 18](#_Toc75898684)

[5.2. Training data and Test data 19](#_Toc75898685)

[5.3 Thuật toán Decision Tree 20](#_Toc75898686)

[5.3.1 Thuật toán phân nhánh information gain 20](#_Toc75898687)

[5.1.2 Thực hiện phân nhánh với chỉ số Gini 24](#_Toc75898688)

[5.4. Thuật toán Naïve Bayes 27](#_Toc75898689)

[5.5 Thuật toán Random Forest 31](#_Toc75898690)

[5.6 Thuật toán Logistic Regression 34](#_Toc75898691)

[5.7. Đánh giá kết quả của các thuật toán được phân tích 34](#_Toc75898692)

# CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI

## Lí do chọn đề tài

## Dữ Liệu

## Dataset Traffic Crashes – Crashes

* Link Dataset :

<https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>

* Thông Tin Về Dữ Liệu :

Dữ liệu về lượng mưa ở các tỉnh thành phố ở Australia trong vòng 10 năm . Những quan sát này được lấy từ hệ thống "thời gian thực" của Cục Khí tượng. Hầu hết dữ liệu được tạo và xử lý tự động. Một số kiểm tra chất lượng đã được thực hiện, nhưng vẫn có thể xuất hiện các giá trị sai sót. Đôi khi, các quan sát sẽ không có sẵn, vì nhiều lý do.

Bảng cho tháng hiện tại được cập nhật hai lần một ngày: một lần vào buổi sáng muộn và một lần nữa vào buổi chiều muộn. Tất cả các quan sát trong tháng phải có sẵn với bản cập nhật buổi sáng vào ngày đầu tiên của tháng tiếp theo. Các tháng trước được cập nhật vài tuần một lần

A picture containing text

Description automatically generated

|  |  |
| --- | --- |
| Số Dòng | Số Cột |
| 145460 | 23 |

* Thông tin của những thuộc tính có trong bộ dữ liệu:

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên Cột** | **Ý Nghĩa** |
| 1 | Date | Ngày và thời gian được nhân viên ghi và nhập vào hệ thống |
| 2 | Location | Vị trí thời tiết tại thời điểm diễn ra được ghi lại. |
| 3 | MinTemp | Nhiệt độ tối thiểu trong 24 giờ đến 9 giờ sáng. Đôi khi chỉ được biết đến ở mức độ gần nhất. |
| 4 | MaxTemp | Nhiệt độ tối đa trong 24 giờ từ 9 giờ sáng. Đôi khi chỉ được biết đến ở mức độ gần nhất. |
| 5 | Rainfall | Lượng mưa trong 24 giờ đến 9 giờ sáng. Được đo đến chính xác đến từng milimet. |
| 6 | Evaporation | Tốc độ bay hơi trong 24 giờ đến 9 giờ sáng |
| 7 | Sunshine | Thời gian nắng trong một ngày. tính bằng giờ |
| 8 | WindGustDir | Hướng gió trong vòng 24 giờ đến nữa đêm |
| 9 | WindGustSpeed | Tốc độ gió trong vòng 24 giờ đến nữa đêm được đo theo km/h. |
| 0 | WindDir9am | Hướng gió trung bình trong 10 phút trước 9 giờ sáng |
| 11 | WindDir3pm | Hướng gió trung bình trong 10 phút trước 3 giờ chiều. |
| 12 | WindSpeed9am | Tốc độ gió trung bình trong 10 phút trước 9 giờ sáng |
| 13 | WindSpeed3pm | Tốc độ gió trung bình trong 10 phút trước 3 giờ chiều. |
| 14 | Humidity9am | Độ ẩm tương đối được đo vào lúc 9 giờ sáng . giá trị % |
| 15 | Humidity3pm | Độ ẩm tương đối được đo vào lúc 3 giờ chiều . giá trị % |
| 16 | Pressure9am | Áp suất khí quyển giảm xuống mực nước biển trung bình lúc 9 giờ sáng |
| 17 | Pressure3pm | Áp suất khí quyển giảm xuống mực nước biển trung bình lúc 3 giờ chiều |
| 18 | Cloud9am | Độ phủ của mây được đo vào lúc 9 giờ sáng . trên thang điểm 8 |
| 19 | Cloud3pm | Độ phủ của mây được đo vào lúc 3 giờ chiều . trên thang điểm 8 |
| 20 | Temp9am | Nhiệt độ vào lúc 9 giờ sáng |
| 21 | Temp3pm | Nhiệt độ vào lúc 3 giờ chiều |
| 22 | RainToday | Ngày hôm nay có mưa không |
| 23 | RainTomorrow | Dự Đoán ngày mai có mưa không được thể hiện bằng giá trị yes hoặc no |

## 2.4 Xóa những thuộc tính không cần thiết :

### 2.4.1 Xóa những thuộc tính trên file không cần thiết :

Table

Description automatically generated

Ta thấy: - Evaporation có 43% dữ liệu thiếu

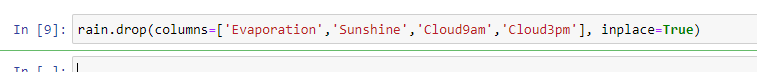
- Sunhine có 48% dữ liệu thiếu

- Clould9am có 38% dữ liệu thiếu

- Clould3pm có 40% dữ liiệu thiếu

Có thể nhận thấy các thuộc tính trên hiện đang thiếu dữ liệu rất nghiêm trọng ,

điều này sẽ gây ảnh hưởng lớn nếu sử dụng thuộc tính này trong quá trình khai thác. Nên mình sẽ xóa các cột trên .



# Chương 3 : Tiền xữ lý dữ liệu

## 3.1 xữ lí dữ liệu còn thiếu

Graphical user interface

Description automatically generated

Ta Thấy những dữ liệu còn trống

\* MinTemp

\* MaxTemp

\* Rainfall

\* WindGustDir

\* WindGustSpeed

\* WindDir9am

\* WindDir3pm

\* WindSpeed9am

\* WindSpeed3pm

\* Humidity9am

\* Humidity3pm

\* Pressure9am

\* Pressure3pm

\* Temp9am

\* Temp3pm

\* RainToday

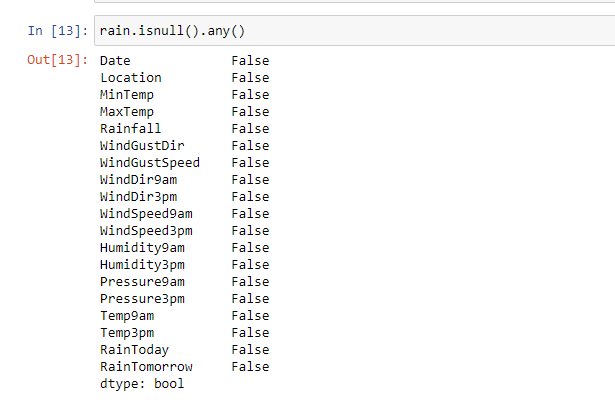
\* RainTomorrow

Text

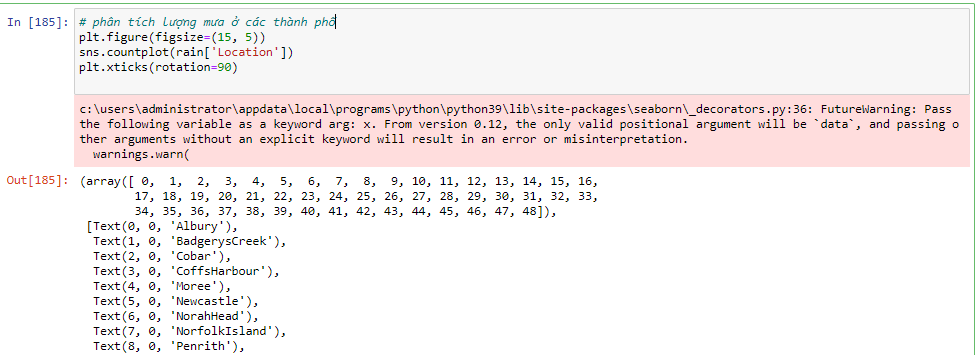
Description automatically generated



## 3.2 Kết quả kiểm tra các giá trị null trong bảng dữ liệu



# Chương 4 : Trực quan hóa dữ liệu



Chart, bar chart, histogram

Description automatically generated

* Sydney là thành phố có lượng mưa cao nhất 2500 mn
* Nhil là thành phố có lượng mưa thấp nhất
* Còn lại lượng mưa các thành phố khác xấp xỉ 2000mn

Chart, bar chart

Description automatically generated

Kết luận : Ta có thể thấy hướng gió Tây là chiếm số lượn nhiều nhất còn các hướng gió còn lại xuất hiện với tần suất ít hơn có thể đoán rắng Australia là 1 nơi rất khô cằn.



Chart, bar chart

Description automatically generated

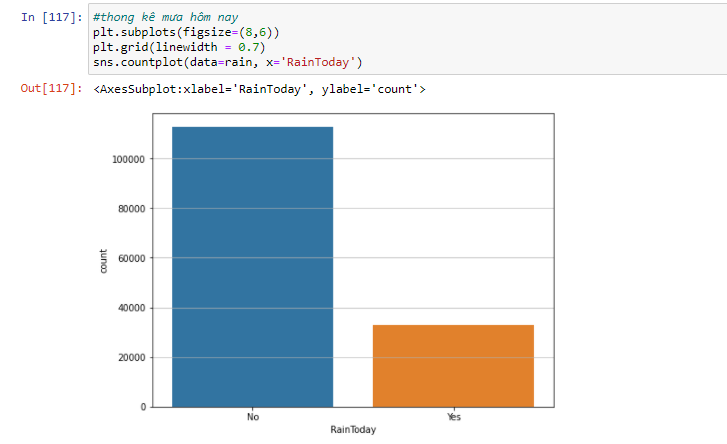
* Hướng gió trong lúc 9h tập trung chủ yếu là hướng gió nam



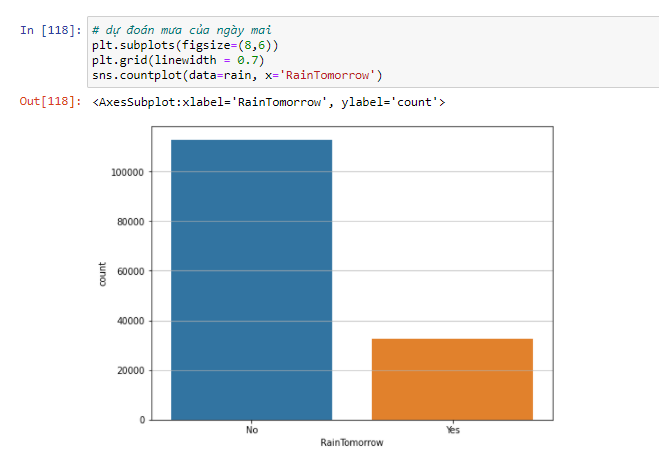
Chart, bar chart

Description automatically generated

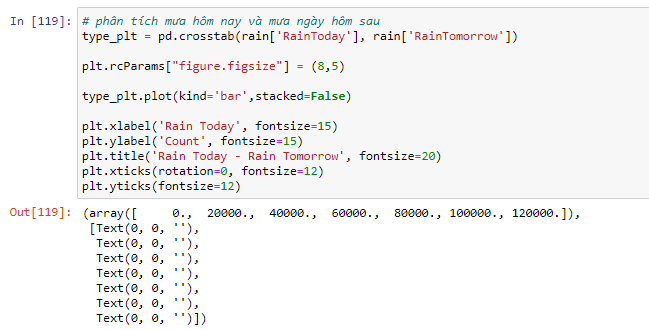
* Hướng gió đông nam là hướng gió có số lần xuất hiện nhiều nhất

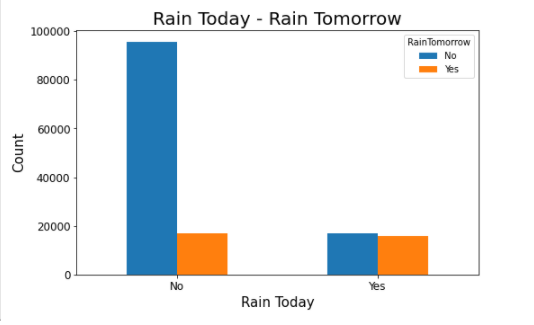


* Khả năng không có mưa cao hơn khả năng có mưa trong ngày



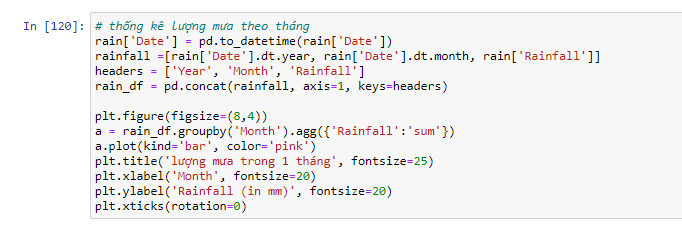
* Dự đoán ngày mai tỉ lệ không mưa cao hơn tỉ lệ có mưa

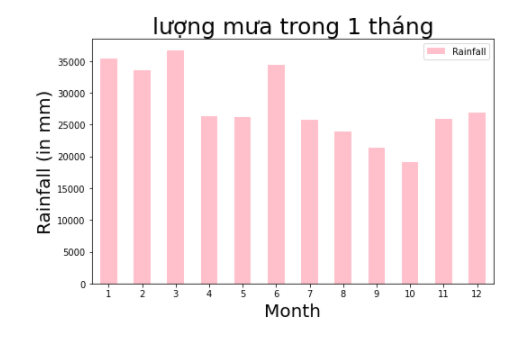




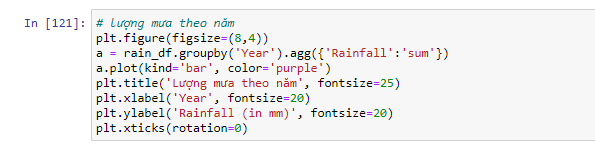
Đây là so sánh mưa ngày hôm nay và ngày mai :

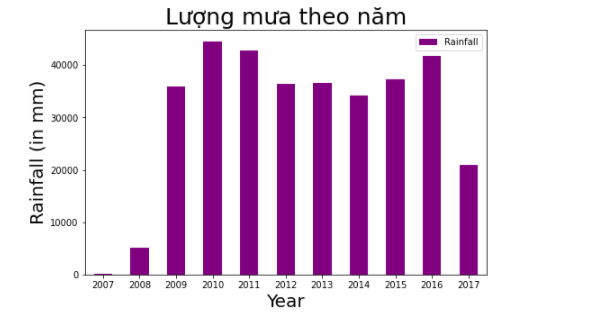
* Với không mưa tỉ lệ nếu ngày hôm nay không có mưa cao hơn nhiều so với hôm sau . nếu ngày hôm nay không có mưa thì khả năng ngày hôm sau có mưa thấp chỉ 25%
* Nếu ngày hôm nay có mưa thì tỉ lệ ngày mai có mưa sẽ khoảng 49%



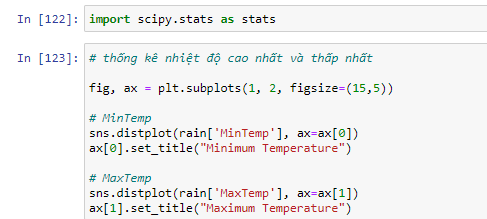


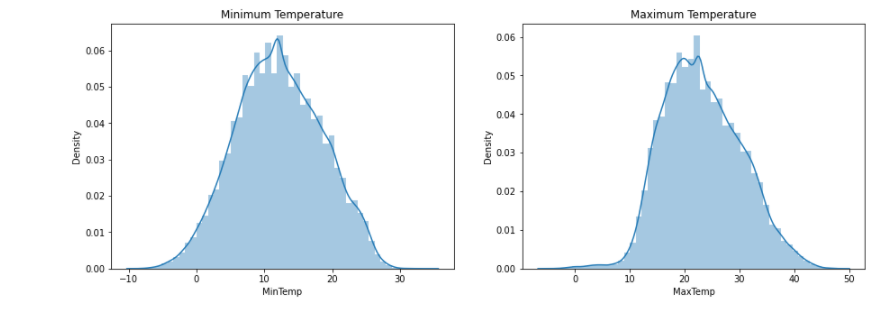
* Lượng mưa cao nhất trong năn sẽ chủ yếu vào tháng 1 , 3 và tháng 6
* Các tháng 8, 9 , 10 là những tháng có lượng mưa ít nhất





* Trong vòng 10 năm thống kê năm 2010 và 2011 là các năm có lượng mưa cao nhât khoảng hơn 45000 mm
* Do thống kê được lập cuối năm 2007 nên số liệu vẫn chưa cụ thể , ta thấy được trên biểu đồ năm 2008 và 2017 có lượng mưa ít nhất dưới 20000 mm.





- Nhiệt độ điểm cao nhất cho nhiệt độ tối thiểu là từ 10 đến 12 độ .

- Nhiệt độ điểm cao nhất cho nhiệt độ tối đa là từ 18 đến 22 độ .

# Chương 5 : chạy thuật toán

## 5.1. Kiểm tra dữ liệu

- xử lí chuyển các dữ liệu dạng object sang dạng int dùng hàm LablEncoder và kết quả sau khi chuyển đổi

A picture containing table

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Kiểu dữ liệu | Yes | No |
| Raintoday | 1 | 0 |
| Raintomorrow | 1 | 0 |

## 5.2. Training data and Test data

Tách cột dữ liệu thành hai phần, một phần chứa các thuộc tính bình thường, một phần chứa riêng thuộc tính quyết định.

+ Các thuộc tính bình thường:

* Date
* Location
* MinTemp
* MaxTemp
* Rainfall
* WindGustDir
* WindGustSpeed
* WindDir9am
* WindDir3pm
* WindSpeed9am
* WindSpeed3pm
* Humidity9am
* Humidity3pm
* Pressure9am
* Pressure3pm
* Temp9am
* Temp3pm
* RainToday

+ Thuộc tính quyết định:

* Raintomorow

Thực hiện tách các cột dữ liệu vào hai biến:

features (chứa các thuộc tính bình thường)

labels (chứa thuộc tính quyết định)

Với rain là DataFrame chứa toàn bộ dữ liệu của lượng mưa ở Australia.

Graphical user interface, application

Description automatically generated

Chia dữ liệu thành 20% test 80% training

## 5.3 Thuật toán Decision Tree

### 5.3.1 Thuật toán phân nhánh information gain

Text

Description automatically generated

Với biến criterion=” entropy”

Ma trận nhầm lẫn

Graphical user interface, text, application

Description automatically generated

* Biểu diễn ma trận lên đồ thị heatmap

Chart, treemap chart

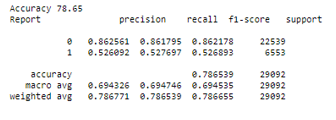
Description automatically generated

* Phân tích thuật toán:

Từ kết quả phân tích ta có ma trận nhầm lẫn sau

|  |  |  |  |
| --- | --- | --- | --- |
| Lớp dự đoán được từ mô hình | | | |
| Lớp trên thực tế |  | Không mưa | Có mưa |
| Không mưa | 19424 | 3095 |
| Có mưa | 3115 | 3458 |

* Phân tích kết quả của ma trận nhầm lẫn:
* TP = 19424 (Với TP (True Positive): kết quả dự báo / phân loại của mô hình là “Positive” đúng với giá trị thực là “Positive”).
* FP = 3115 (Với FP (False Positive)): kết quả dự báo / phân loại của mô hình là “Positive” sai với giá trị thực là “Negative”) hay còn gọi là sai lầm loại I (Type I Error).
* FN = 3095 (Với FN (False Negative)): kết quả dự báo / phân loại mô hình là “Negative” sai với giá trị thực tế là “Positive”) hay còn gọi là sai lầm loại II (Type II Error).
* TN = 3458 (Với TN (True Negative)): kết quả dự báo / phân loại mô hình là “Negative” đúng với giá trị thực tế là “Negative”).
* Các chỉ số (metrics):



Precision (Positive Predicted Value) là tỷ lệ các trường hợp Positive của mô hình dự báo được phân loại đúng.

Sensitivity (Recall) là tỷ lệ các trường hợp Positive thực sự (actual) được phân loại đúng.

Negative Predictive Value là tỷ lệ các trường hợp Negative được phân loại đúng.

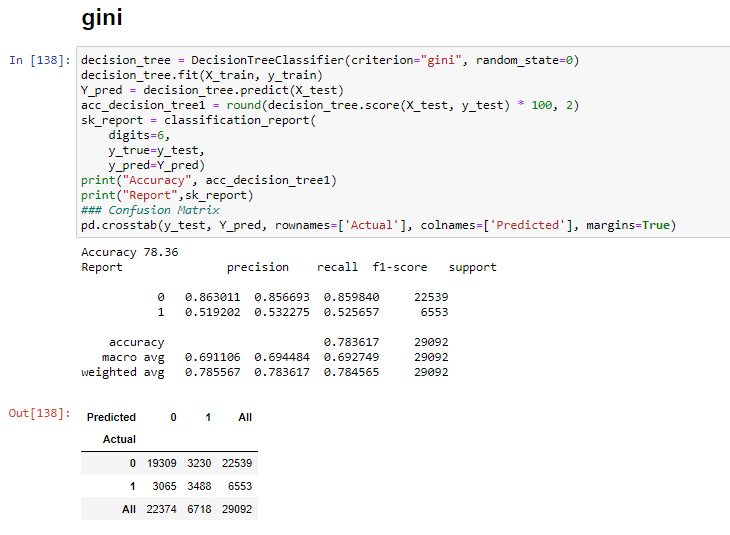
Specificity là tỷ lệ các trường hợp Negative thực sự (actual) đã được phân loại đúng.

Accuracy Rate là tỷ lệ dự báo đúng cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

Error Rate bằng 1 - Accuracy Rate là tỷ lệ dự báo sai cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không).

### 5.1.2 Thực hiện phân nhánh với chỉ số Gini



* Phân tích thuật toán:

Từ kết quả phân tích ta có ma trận nhầm lẫn sau

|  |  |  |  |
| --- | --- | --- | --- |
| Lớp dự đoán được từ mô hình | | | |
| Lớp trên thực tế |  | Không mưa | Có mưa |
| Không mưa | 19309 | 3320 |
| Có mưa | 3065 | 3348 |

* Phân tích kết quả của ma trận nhầm lẫn:
* TP = 19309 (Với TP (True Positive): kết quả dự báo / phân loại của mô hình là “Positive” đúng với giá trị thực là “Positive”).
* FP = 3320(Với FP (False Positive)): kết quả dự báo / phân loại của mô hình là “Positive” sai với giá trị thực là “Negative”) hay còn gọi là sai lầm loại I (Type I Error).
* FN = 3065 (Với FN (False Negative)): kết quả dự báo / phân loại mô hình là “Negative” sai với giá trị thực tế là “Positive”) hay còn gọi là sai lầm loại II (Type II Error).
* TN = 3348 (Với TN (True Negative)): kết quả dự báo / phân loại mô hình là “Negative” đúng với giá trị thực tế là “Negative”).
* Các chỉ số (metrics):

Text, letter

Description automatically generated

Precision (Positive Predicted Value) là tỷ lệ các trường hợp Positive của mô hình dự báo được phân loại đúng.

Sensitivity (Recall) là tỷ lệ các trường hợp Positive thực sự (actual) được phân loại đúng.

Negative Predictive Value là tỷ lệ các trường hợp Negative được phân loại đúng.

Specificity là tỷ lệ các trường hợp Negative thực sự (actual) đã được phân loại đúng.

Accuracy Rate là tỷ lệ dự báo đúng cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

Error Rate bằng 1 - Accuracy Rate là tỷ lệ dự báo sai cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không).

## 5.4. Thuật toán Naïve Bayes

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Ma trận nhầm lẫn

Graphical user interface, text

Description automatically generated

* Biểu diễn ma trận lên đồ thị heatmap

Chart, waterfall chart, treemap chart

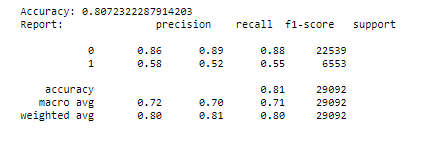
Description automatically generated

* Phân tích thuật toán:

Từ kết quả phân tích ta có ma trận nhầm lẫn sau

|  |  |  |  |
| --- | --- | --- | --- |
| Lớp dự đoán được từ mô hình | | | |
| Lớp trên thực tế |  | Không mưa | Có mưa |
| Không mưa | 20105 | 2434 |
| Có mưa | 3174 | 3379 |

* Phân tích kết quả của ma trận nhầm lẫn:
* TP = 19424 (Với TP (True Positive): kết quả dự báo / phân loại của mô hình là “Positive” đúng với giá trị thực là “Positive”).
* FP = 3115 (Với FP (False Positive)): kết quả dự báo / phân loại của mô hình là “Positive” sai với giá trị thực là “Negative”) hay còn gọi là sai lầm loại I (Type I Error).
* FN = 3095 (Với FN (False Negative)): kết quả dự báo / phân loại mô hình là “Negative” sai với giá trị thực tế là “Positive”) hay còn gọi là sai lầm loại II (Type II Error).
* TN = 3458 (Với TN (True Negative)): kết quả dự báo / phân loại mô hình là “Negative” đúng với giá trị thực tế là “Negative”).
* Các chỉ số (metrics):



Precision (Positive Predicted Value) là tỷ lệ các trường hợp Positive của mô hình dự báo được phân loại đúng.

Sensitivity (Recall) là tỷ lệ các trường hợp Positive thực sự (actual) được phân loại đúng.

Negative Predictive Value là tỷ lệ các trường hợp Negative được phân loại đúng.

Specificity là tỷ lệ các trường hợp Negative thực sự (actual) đã được phân loại đúng.

Accuracy Rate là tỷ lệ dự báo đúng cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

Error Rate bằng 1 - Accuracy Rate là tỷ lệ dự báo sai cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không).

## 5.5 Thuật toán Random Forest

Graphical user interface, text, application, chat or text message

Description automatically generated

Table

Description automatically generated

* Phân tích thuật toán:

Từ kết quả phân tích ta có ma trận nhầm lẫn sau

|  |  |  |  |
| --- | --- | --- | --- |
| Lớp dự đoán được từ mô hình | | | |
| Lớp trên thực tế |  | Không mưa | Có mưa |
| Không mưa | 21480 | 1059 |
| Có mưa | 3356 | 3197 |

* Phân tích kết quả của ma trận nhầm lẫn:
* TP = 21480 (Với TP (True Positive): kết quả dự báo / phân loại của mô hình là “Positive” đúng với giá trị thực là “Positive”).
* FP = 1059 (Với FP (False Positive)): kết quả dự báo / phân loại của mô hình là “Positive” sai với giá trị thực là “Negative”) hay còn gọi là sai lầm loại I (Type I Error).
* FN = 3356 (Với FN (False Negative)): kết quả dự báo / phân loại mô hình là “Negative” sai với giá trị thực tế là “Positive”) hay còn gọi là sai lầm loại II (Type II Error).
* TN = 3197 (Với TN (True Negative)): kết quả dự báo / phân loại mô hình là “Negative” đúng với giá trị thực tế là “Negative”).
* Các chỉ số (metrics):

Text

Description automatically generated

Precision (Positive Predicted Value) là tỷ lệ các trường hợp Positive của mô hình dự báo được phân loại đúng.

Sensitivity (Recall) là tỷ lệ các trường hợp Positive thực sự (actual) được phân loại đúng.

Negative Predictive Value là tỷ lệ các trường hợp Negative được phân loại đúng.

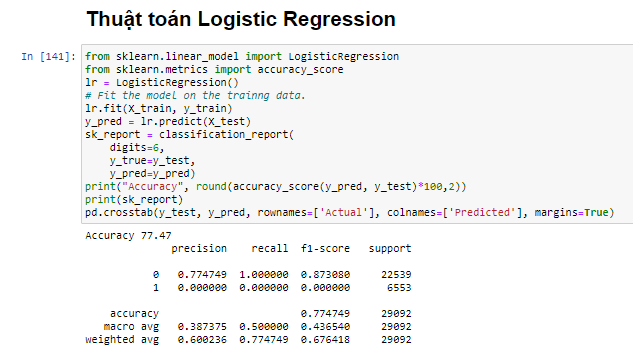
Specificity là tỷ lệ các trường hợp Negative thực sự (actual) đã được phân loại đúng.

Accuracy Rate là tỷ lệ dự báo đúng cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

Error Rate bằng 1 - Accuracy Rate là tỷ lệ dự báo sai cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không).

## 5.6 Thuật toán Logistic Regression



Text

Description automatically generated

## 5.7. Đánh giá kết quả của các thuật toán được phân tích

Ta có giá trị bằng 1 của biến mục tiêu nghĩa là có khả năng có mưa trong ngày mai. Gía trị bằng 0 của biến mục tiêu nghĩa là có khả năng không có mưa trong ngày mai

Bởi vì bài toán đặt ra là xác định đự đoán mưa của ngày hôm sau. Vì vậy chúng ta quan tâm đến khả năng có mưa.

Do đó giá trị bằng 1 sẽ được quan tâm. Vì nó là khả năng có mưa của ngày hôm sau.

Vì vậy giá trị Recall và Specificity đóng vai trò quan trọng trong trong việc chọn ra thuật toán nào sẽ dự đoán chính xác nhất bài toán được đặt ra.

Hai giá trị Recall (*Negative Predictive Value là tỷ lệ các trường hợp Negative được phân loại đúng*) và Specificity (*Specificity là tỷ lệ các trường hợp Negative thực sự (actual) đã được phân loại đúng*) của thuật toán nào lớn nhất thì tức là thuật toán đó sẽ đưa ra dự báo tốt nhất cho bài toán được đặt ra.

|  |  |  |
| --- | --- | --- |
|  | ***Negative Predictive Value*** | ***Specificity*** |
| **Decision Tree (entropy)** | 0,526 | 0,527 |
| **Decision Tree (gini)** | 0,51 | 0,53 |
| **Naive Bayes** | 0,58 | 0,52 |
| **Random forest** | 0,75 | 0,487 |
| **Logistic Regression** | 0 | 0 |

Như chúng em đã thực hiện các thuật toán Decision Tree, Naive Bayes, Logistic Regression, Random Forest để dự đoán mưa ngày hôm sau. Có hai điều mà chúng em có thể kết luận:

* Ta thấy được cả hai thuật toán Decision Tree (entropy) và Random Forest dự đoán tốt hơn nhiều về mặt dự đoán .
* Ta sử dụng tính năng quan trọng và có các yếu tố sao gây ra ảnh hưởng nhiều nhất đến mưa của ngày hôm sau là:

Graphical user interface, application

Description automatically generated

Từ mô hình cho ta thấy được hai yếu tố quan trọng gây ra mưa là Humdity3pm và Pressure3pm.