**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG CÔNG NGHIỆP THỰC PHẨM**

**2020-2021**

**🙞🙜**

****

Khai Thác Dữ Liệu

ĐỀ TÀI BÁO CÁO :

Dự báo thời tiết

Giáo Viên Hướng Dẫn :Ngô Dương Hà

Sinh Viên Thực Hiện: 2001181383 Vũ Minh Trung

2001180390 Nguyễn Đức Trung

**MỤC LỤC**

[I. Dữ Liệu : 2](#_Toc59995217)

[II. Xử Lý Dữ Liệu: 4](#_Toc59995218)

[III. Cây Quyết Định: 9](#_Toc59995219)

[1. **Giới thiệu về cây quyết định:** 9](#_Toc59995220)

[2. **Làm sao để sử dụng cây quyết định trong dự đoán lớp của các dữ liệu chưa biết** 11](#_Toc59995221)

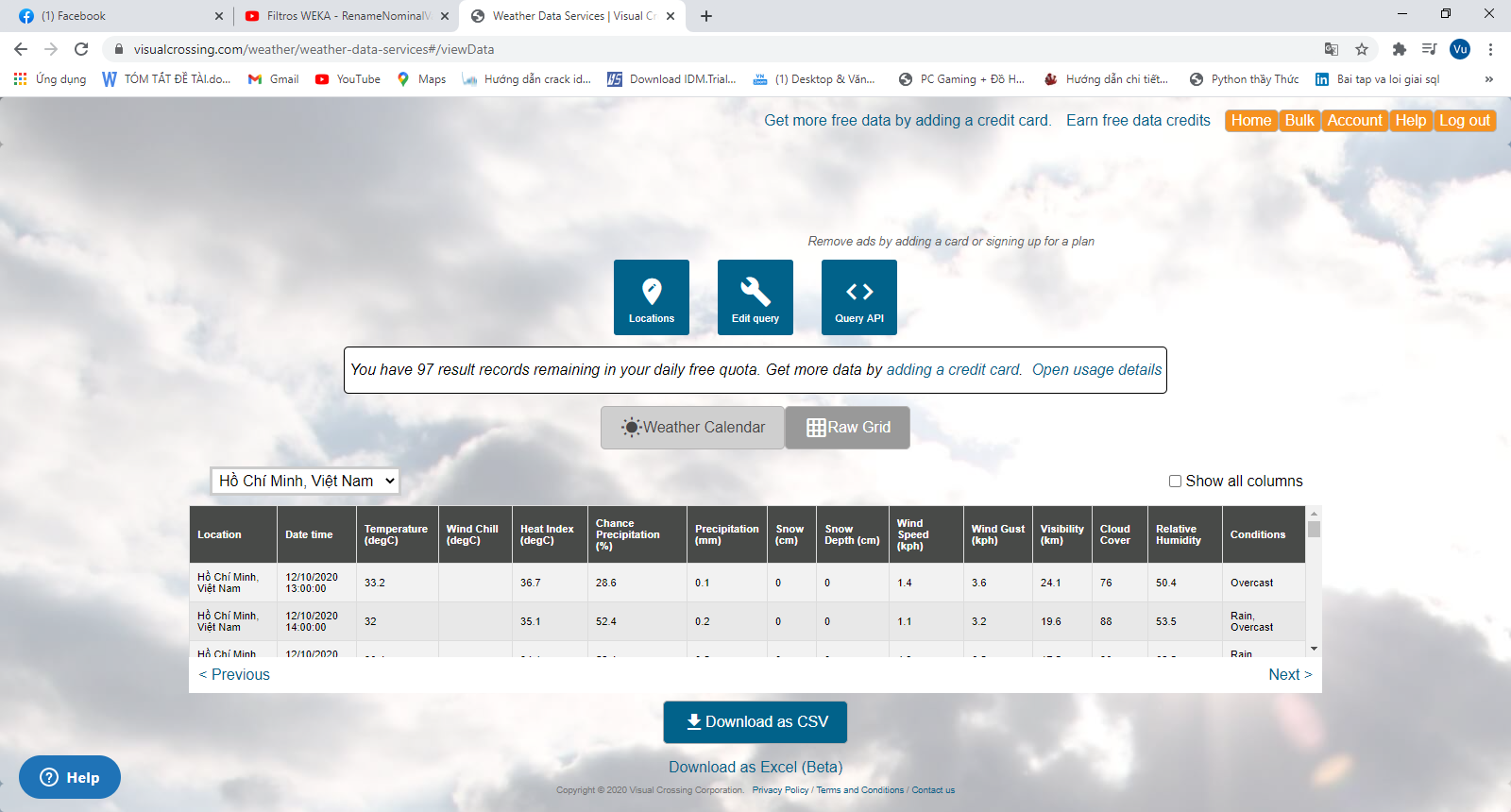
[3. Nhận xét: 12](#_Toc59995222)

[IV. Đánh Giá: 13](#_Toc59995223)

[V. Phân công: 13](#_Toc59995224)

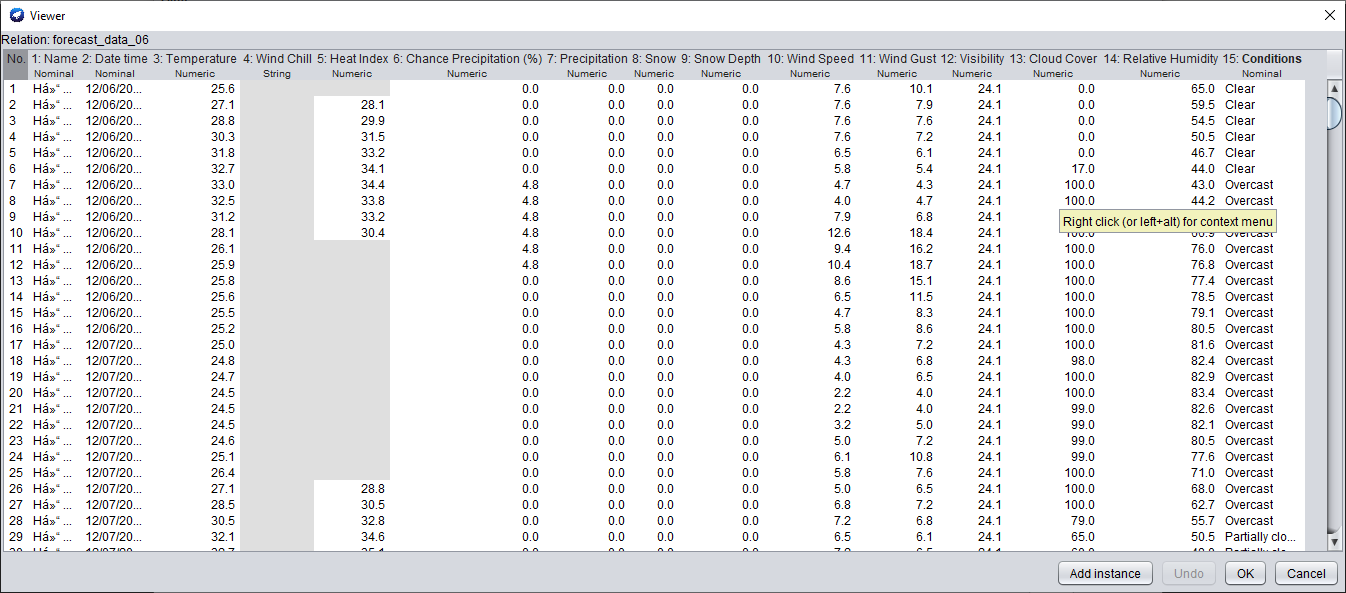
1. Dữ Liệu :

* Vào trang web VISUAL CROSSING
* Lấy dữ liệu theo giờ trong ngày về:



Khi Lấy Về Dữ Liệu sẽ có những thông tin như thế này:

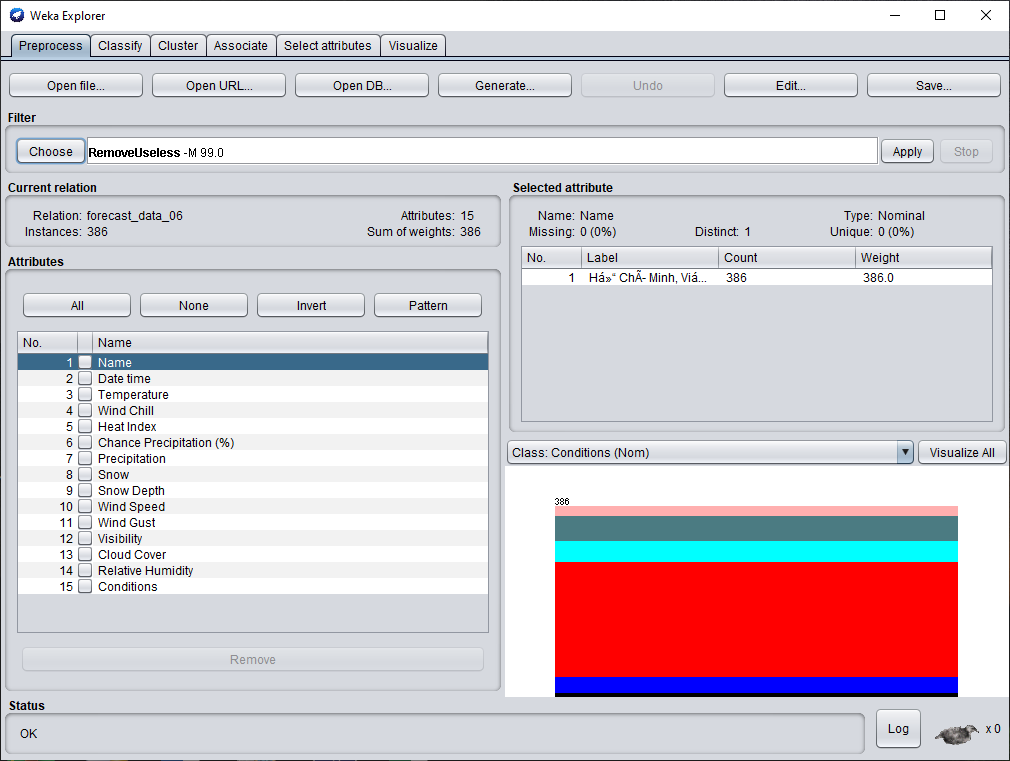
* Dùng WEKA để hiển thị



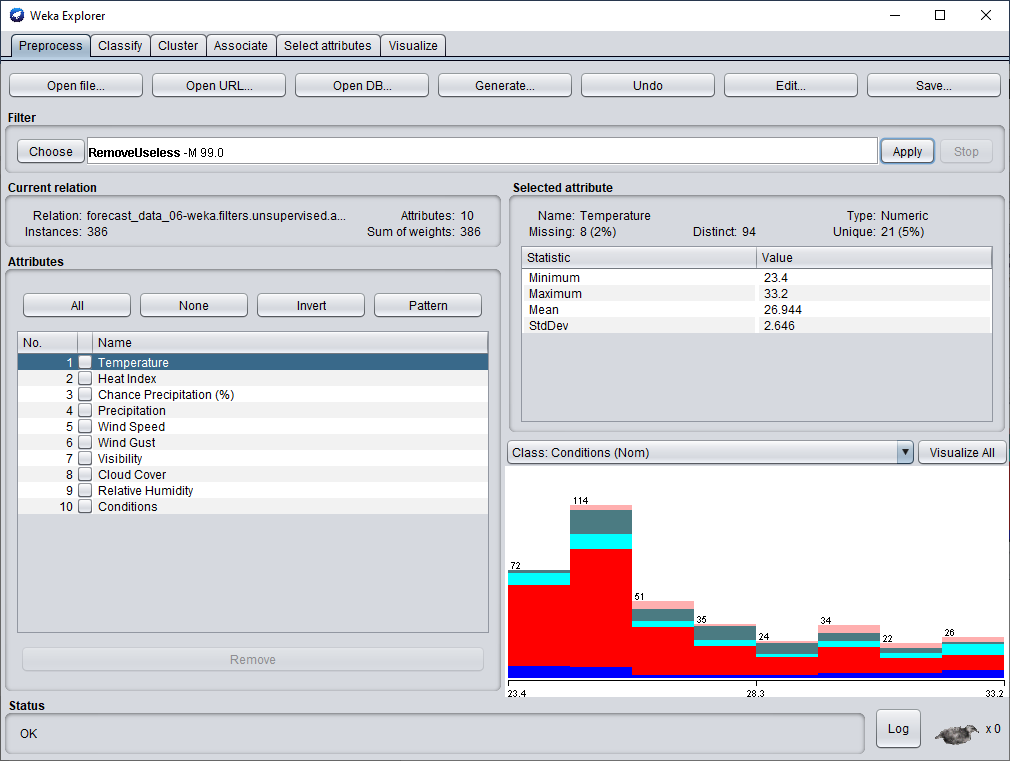
* Trên là những dữ liệu thô chưa được xử lý.

1. Xử Lý Dữ Liệu:

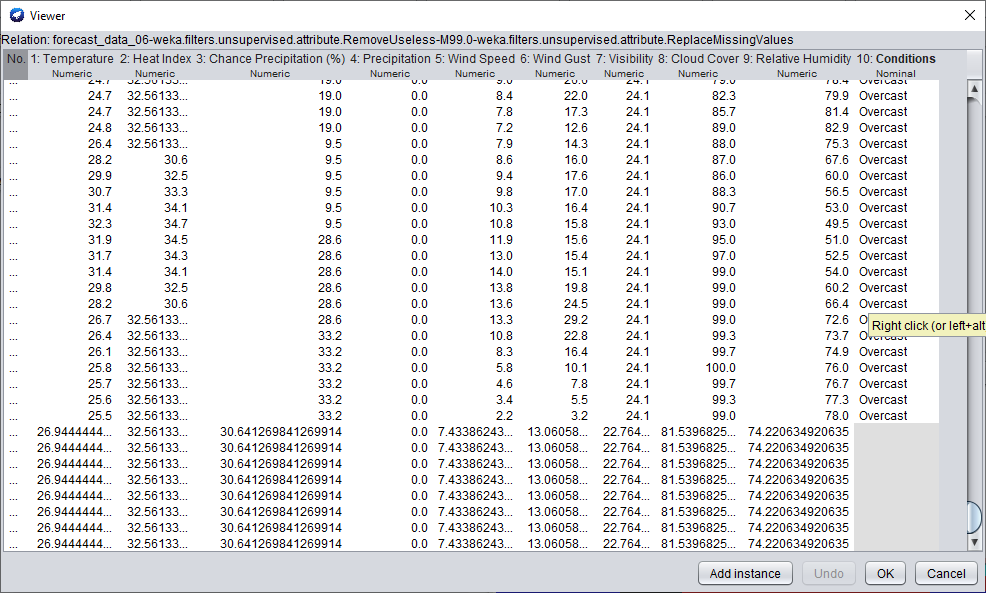
* Dùng weka sử dụng kĩ thuật học không giám sát :remove useless

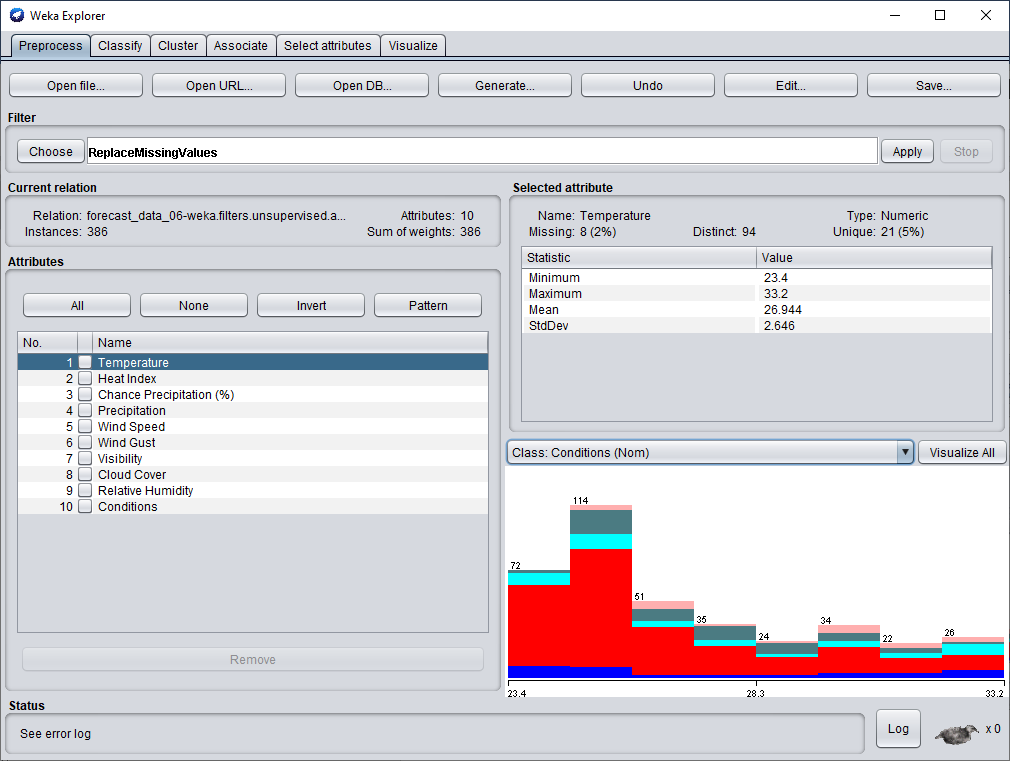


Kết quả:

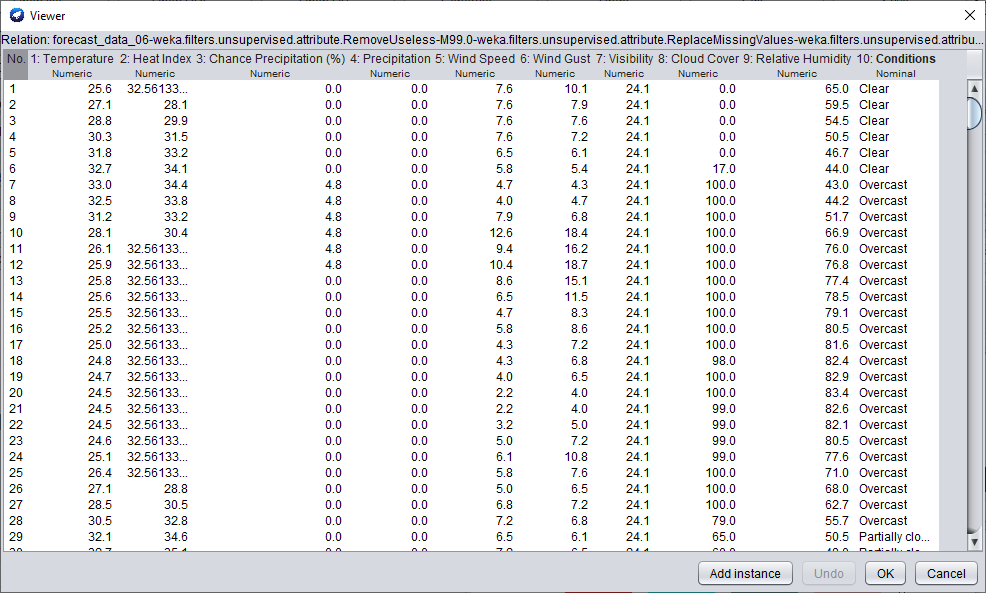


Tuy nhiên vẫn còn vài dữ liệu bị mất

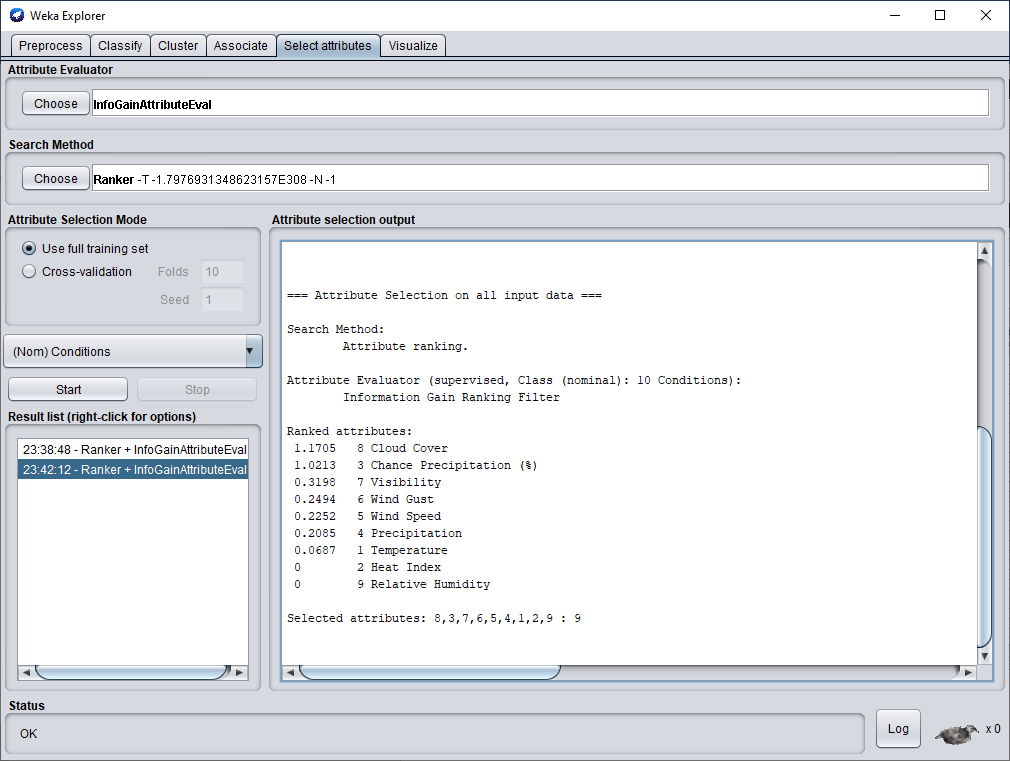




Kết quả :

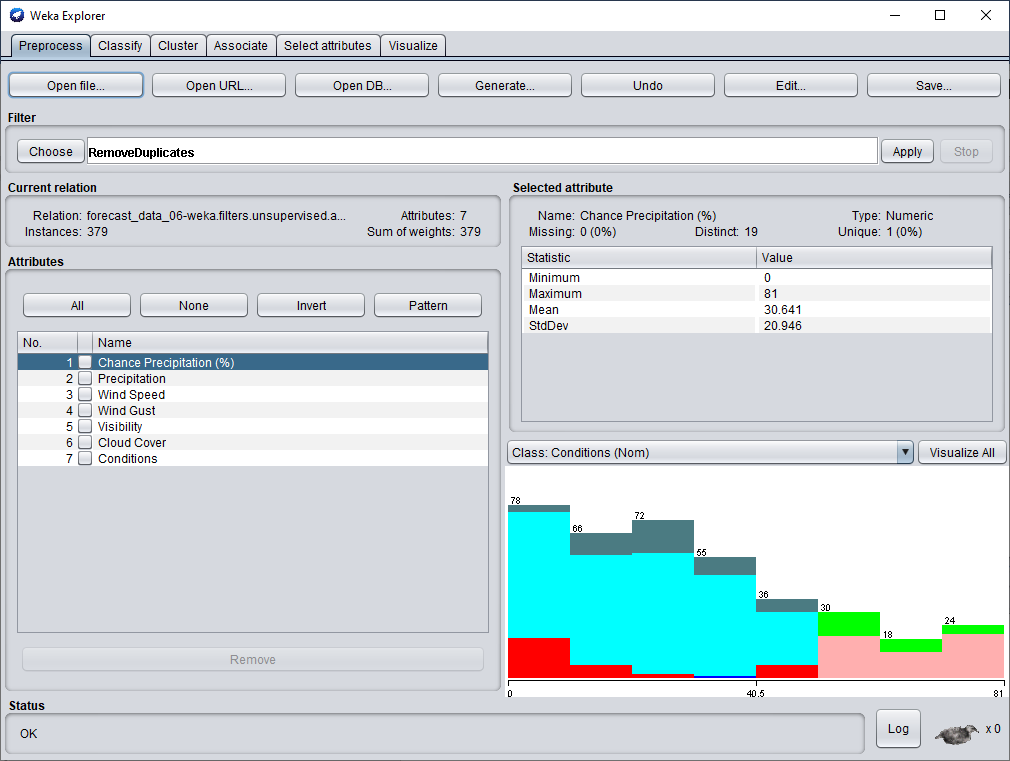


* Sau khi đã xử lí các dữ liệu bị mất
* Dùng thuật toán

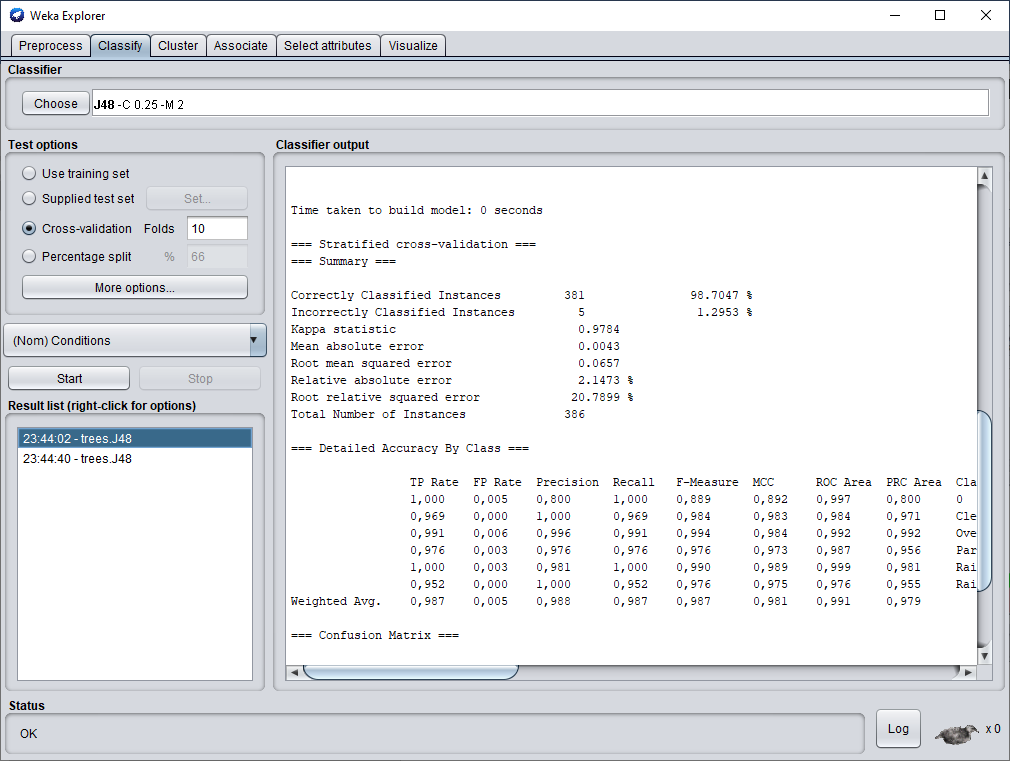


Để cần chọn thuộc tính nào để độ giảm Entropy là thấp nhất.

* Thì thấy 3 dòng cuối có kết quả =0
* Xóa dữ liệu 1,2,9;



Có thể thấy dữ liệu đã được làm sạch



1. Cây Quyết Định:
2. **Giới thiệu về cây quyết định:**

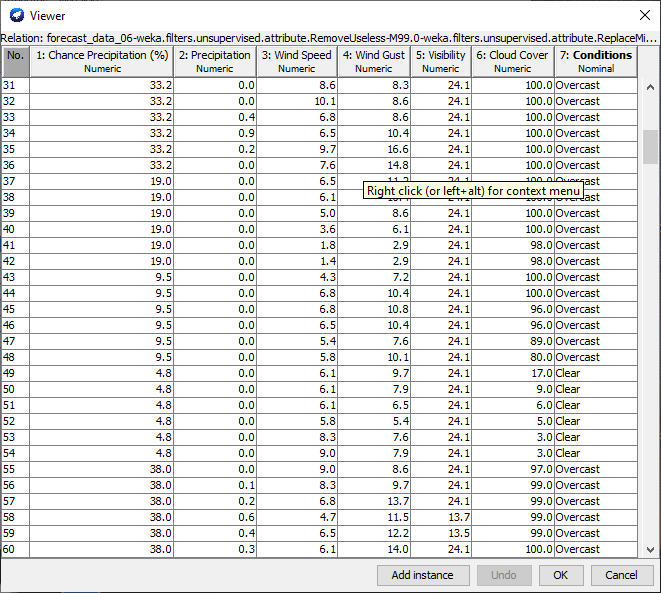
-Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật (series of rules). Các thuộc tính của đối tượng (ngoại trừ thuộc tính phân lớp – Category attribute) có thể thuộc các kiểu dữ liệu khác nhau (Binary, Nominal, ordinal, quantitative values) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.

-Tóm lại, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết (unseen data)

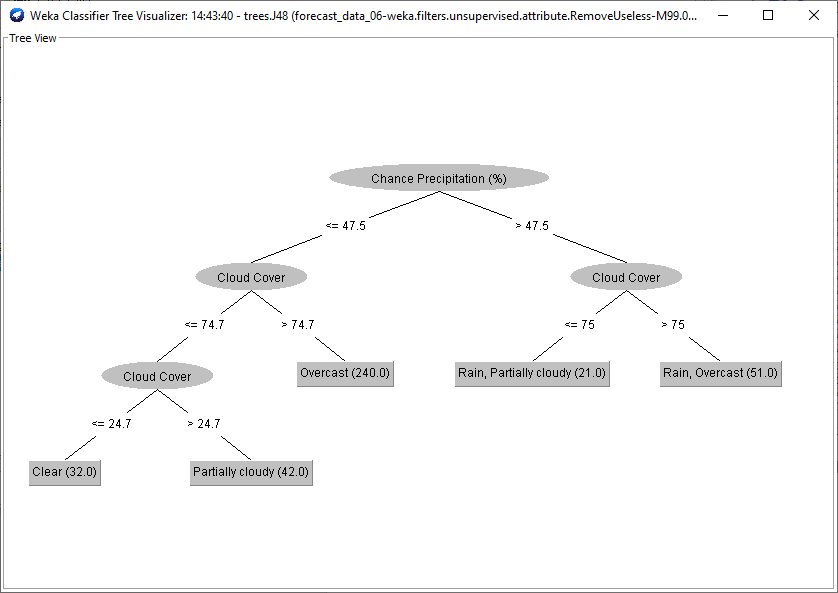
Để  giúp các bạn ứng dụng được cây quyết định trong bài toán phân lớp chúng ta cùng bắt đầu với ví dụ dưới đây:

Ta có dữ liệu (training data) trên. Mỗi đối tượng được mô tả bởi 7 thuộc tính là **Chance\_precipitation,Precipitation,WindSpeed,Wind Gust,Visibility,Cloud\_cover** và 1 thuộc tính phân loại **Conditions**

|  |  |
| --- | --- |
| Attributes | Classes |



Dựa vào Training Data ở trên, chúng ta có thể tạo ra cây quyết định như sau



1. **Làm sao để sử dụng cây quyết định trong dự đoán lớp của các dữ liệu chưa biết**

Mục đích chính của cây quyết định là dùng để dự đoán lớp (xác định lớp) của các đối tượng chưa biết (unseen data). Giả sử rằng ta có dữ liệu về 3 các giá trị dữ liệu đã biết về các thuộc tính **Chance\_precipitation,Precipitation,WindSpeed,WindGust,Visibility,Cloud\_cover**. Tuy nhiên ta chưa biết nó sẽ rơi vào trong Condition nào (Clear,Overcast,Partially cloud,,). Nhiệm vụ của chúng ta là sử dụng cây quyết định đã tạo ra để dự đoán (predict) 3 mẫu sẽ rơi vào Conditions nào dựa vào 4 thuộc tính của họ. Dữ liệu dưới đây còn được gọi là Testing Data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Chance\_precipitation** | **Precipitation** | **WindSpeed** | **Wind Gust** | **Visibility** | **Cloud\_cover** | Conditions |
| 0.0 | 2.1 | 85.0 | 25.0 | 13.2 | 0.23 | ? |
| 2.3 | 2.1 | 85.0 | 25.0 | 13.2 | 0.23 | ? |
| 4.5 | 01.2 | 85.0 | 25.0 | 13.2 | 0.23 | ? |

**Theo cây quyết định trên, các luật (Series of Rules) được sinh ra từ cây quyết định dùng để dự đoán như sau:**

Rule 1 : If Chance\_precipitation > 47.5 and Cloud cover > 75 then Rain+Overcast

Rule2: If Chance\_precipitation > 47.5 and Cloud cover <= 75 then Rain+PartitalCloud

Rule3: If Chance\_precipitation <=47.5 and Cloud cover > 74.7 then Overcast

Rule4: If Chance\_precipitation <=47.5 and Cloud cover > 24.7 then PartitalCloud

Rule 5: If Chance\_precipitation <=47.5 and Cloud cover < 24.7 then Clear

1. Nhận xét:

* Cây quyết định là một phương pháp phân lớp rất hiệu quả và dễ hiểu. Tuy nhiên có một số chú ý khi sử dụng cây quyết định trong xây dựng các mô hình phân lớp như sau:
* Hiệu của phân lớp của cây quyết định (Series of Rules) phụ thuộc rất lớn vào training data. Chẳn hạn cây quyết định được tạo ra bởi chỉ giới hạn 10 samples training data trong ví dụ trên thì hiệu quả ứng dụng cây quyết định để dự đoán các trường hợp khác là không cao (thường training data phải đủ lớn và tin cậy) và vì vậy ta không thể nói rằng tập các luật (Series of Rules) được sinh ra bởi cây quyết định trên là tập luật tốt nhất.
* Có rất nhiều thuật toán phân lớp như ID3, J48, C4.5, CART (Classification and Regression Tree),… Việc chọn thuật toán nào để có hiệu quả phân lớp cao tuy thuộc vào rất nhiều yếu tố, trong đó cấu trúc dữ liệu ảnh hưởng rất lớn đến kết quả của các thuật toán. Chẳn hạn như thuật toán ID3 và CART cho hiệu quả phân lớp rất cao đối với các trường dữ liệu số (quantitative value) trong khi đó các thuật toán như J48, C4.5 có hiệu quả hơn đối với các dữ liệu Qualititive value (ordinal, Binary, nominal).

1. Đánh Giá:

Trong các thuật toán xây dựng cây quyết định (ID3, J48, C4.5,...) thường dùng Entropy (Information Gain - độ lợi thông tin) để xác định Root Node của cây. Ngoài ra, có thể sử dụng một số cách khác như Bayesian with K2 Prior, hay Bayesian Dirichlet Equivalent (BDE) Prior để chọn Root Node cho cây.

Thuật toán xây dựng Decision Tree của Microsoft là thuật toán lai ghép giữ nhiều thuật toán và có thể chọn phương thức xác định Root Node của Decision Tree bằng cách thay đổi tham số SCORE\_METHOD của thuật toán

Nếu sử dụng ID3 để tạo cây thì muốn có cây quyết định tối ưu ( cây quyết định thực chất là tập các luật suy diễn) thì phải qua quá trình tinh chỉnh, cắt tỉa cây (Pruning tree)

1. Phân công:

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ Và Tên** | **Tiến Độ** |
| 2001181383 | Vũ Minh Trung | 80% |
| 2001180390 | Nguyễn Đức Trung | 20% |