ĐỒ ÁN CUỐI KỲ

DỰ BÁO THỜI TIẾT

THÀNH VIÊN

THÁI HOÀNG HUY – 18127109
TRẦN XUÂN LỘC – 18127131

NỘI DUNG

- KHAI THÁC DỮ LIỆU
- KHÁM PHÁ DỮ LIỆU
- ĐẶT CÂU HỎI
- ĐẶT VẤN ĐỀ
- BỔ SUNG DỮ LIỆU
- TIỀN XỬ LÝ VÀ KHÁM PHÁ DỮ LIỆU TRÊN TẬP HUẤN LUYỆN
- XÂY DỰNG MÔ HÌNH
- THỬ NGHIỆM MÔ HÌNH
- Ý TƯỞNG CẢI TIẾN

KHAI THÁC DỮ LIỆU

NGUỒN DỮ LIỆU

- Dữ liệu sử dụng trong đồ án lần này được lấy từ api của trang web Ambee Weather.
- Giới hạn lượt truy cập cho mỗi tài khoản (500 lượt/ngày).
- Dữ liệu tại một số địa điểm có thể không được cung cấp đầy đủ bởi trang web đó.

CÁCH THU THẬP

- Nhận thông tin tọa độ kinh độ (longitude) , vĩ độ (latitude) và key thể hiện cho khóa API mà trang web cung cấp cho tài khoản của ta trong quá trình đăng kí tài khoản.
- Thu thập dữ liệu 13 ngày đầu tiên trong mỗi tháng vào 3 khoảng thời gian trong ngày từ 7-9h sáng, 11-13h chiều và 16-18h tối trong năm 2020.
- Khoảng thời gian cách nhau mỗi 2 tiếng.

BIẾN ĐỔI DỮ LIỆU

DỮ LIỆU THÔ

{"MESSAGE":"SUCCESS","DATA":
{"LAT":35.86166,"LNG":104.195397,"HISTORY":
[{"TIME":1577865600,"SUMMARY":"CLEAR","ICON":"CLEAR-
DAY","TEMPERATURE":39.14,"APPARENTTEMPERATURE":37.21,"DE
WPOINT":19.89,"HUMIDITY":0.46,"PRESSURE":1016.7,"WINDSPEE
D":3.16,"WINDGUST":4.08,"WINDBEARING":100,"CLOUDCOVER":0.
06,"UVINDEX":1,"VISIBILITY":10,"OZONE":303.9}]}}

DỮ LIỆU THU ĐƯỢC

MESSAGE,LAT,LNG,TIME,SUMMARY,ICON,TEMPERATURE,APPARENT
TEMPERATURE,DEWPOINT,HUMIDITY,PRESSURE,WINDSPEED,WIND
GUST,WINDBEARING,CLOUDCOVER,UVINDEX,VISIBILITY,OZONE,CO
UNTRY
SUCCESS,35.86166,104.195397,1577865600,CLEAR,CLEAR-
DAY,39.14,37.21,19.89,0.46,1016.7,3.16,4.08,100,0.06,1,10,303.9

KHÁM PHÁ DỮ LIỆU

DỮ LIỆU NHÌN BẰNG MẮT THƯỜNG

message	lat	lng	time	summary	icon	temperature	apparentTemperature	dewPoint	humidity	pressure	windSpeed	windGust	windBearing	cloudCover	uvIndex	visibility	ozone	Country
success	-38.416097	-63.616672	1577865600	Mostly Cloudy	partly-cloudy-night	63.37	63.47	59.27	0.86	1016.9	5.58	8.97	138	0.76	0	10	258.8	Argentina
success	-38.416097	-63.616672	1577880000	Mostly Cloudy	partly-cloudy-day	67.29	67.41	60.4	0.79	1019.2	9.64	14.35	130	0.87	3	10	259.4	Argentina
success	-38.416097	-63.616672	1577898000	Partly Cloudy	partly-cloudy-day	83.6	83.6	57.2	0.41	1016.9	13.11	13.33	112	0.55	9	10	260.2	Argentina
success	-38.416097	-63.616672	1577952000	Clear	clear-night	61.84	61.84	56.23	0.82	1017.2	6.67	18.71	27	0	0	10	260.5	Argentina
success	-38.416097	-63.616672	1577966400	Clear	clear-day	72.05	72.05	59.58	0.65	1018.5	11.66	14.93	9	0.25	3	10	263.6	Argentina
success	-38.416097	-63.616672	1577984400	Partly Cloudy	partly-cloudy-day	86.86	86.86	47.85	0.26	1016.7	10.09	11.62	29	0.48	9	10	266.5	Argentina
success	-38.416097	-63.616672	1578038400	Clear	clear-night	67.09	67.09	59.05	0.75	1017.8	9.89	25.67	13	0	0	10	264.5	Argentina
success	-38.416097	-63.616672	1578052800	Clear	clear-day	74.52	74.81	62.78	0.67	1018.9	15.65	21.42	4	0	3	10	262.2	Argentina
success	-38.416097	-63.616672	1578070800	Clear	clear-day	90.71	90.71	52.95	0.28	1015.1	14.09	15.9	4	0	13	10	257.7	Argentina
success	-38.416097	-63.616672	1578124800	Mostly Cloudy	partly-cloudy-night	74.36	74.44	60.84	0.63	1011.4	15.3	31	333	0.7	0	10	254.6	Argentina
success	-38.416097	-63.616672	1578139200	Overcast	cloudy	77.12	77.48	63.46	0.63	1011.6	11.04	23.87	352	1	3	10	254.9	Argentina
success	-38.416097	-63.616672	1578157200	Clear	clear-day	87.72	87.72	59.64	0.39	1009.2	13.18	15.37	150	0.31	11	10	256.1	Argentina
success	-38.416097	-63.616672	1578211200	Clear	clear-night	61.02	61.16	58.78	0.92	1006.5	10.14	18.16	118	0.02	0	10	262.2	Argentina
success	-38.416097	-63.616672	1578225600	Clear	clear-day	69.62	70.06	63.29	0.8	1005.5	9.74	11.96	69	0.23	3	10	263	Argentina
success	-38.416097	-63.616672	1578243600	Overcast	cloudy	85.08	85.27	60.95	0.44	1000.3	9.8	11.7	195	1	5	10	275.6	Argentina
success	-38.416097	-63.616672	1578297600	Clear	clear-night	58.78	58.78	48.93	0.7	1003.5	9.11	20.72	173	0.23	0	10	308.4	Argentina
success	-38.416097	-63.616672	1578312000	Clear	clear-day	66.59	66.59	42.13	0.41	1005.3	16.88	22.21	227	0	3	10	297.8	Argentina
success	-38.416097	-63.616672	1578330000	Clear	clear-day	75.46	75.46	26.46	0.16	1005	17.57	24.18	230	0.22	10	10	297.9	Argentina
success	-38.416097	-63.616672	1578384000	Clear	clear-night	60.79	60.79	31.54	0.33	1004.6	8.38	18.79	313	0.13	0	10	285.8	Argentina
success	-38.416097	-63.616672	1578398400	Clear	clear-day	71.81	71.81	36.94	0.28	1006.1	14.06	16.37	263	0	3	10	290.5	Argentina
success	-38.416097	-63.616672	1578416400	Clear	clear-day	85.51	85.51	32.74	0.15	1003.2	17.38	25.63	289	0	11	10	289.5	Argentina
success	-38.416097	-63.616672	1578470400	Clear	clear-night	66.76	66.76	52.95	0.61	1001.6	4.79	6.49	335	0.21	0	10	283	Argentina
success	-38.416097	-63.616672	1578484800	Mostly Cloudy	partly-cloudy-day	74.41	74.41	52.33	0.46	1002.7	5.1	9.66	161	0.77	2	10	283.2	Argentina
success	-38.416097	-63.616672	1578502800	Overcast	cloudy	83.09	83.09	47.78	0.29	1000.6	7.18	7.5	34	0.98	5	10	284.5	Argentina

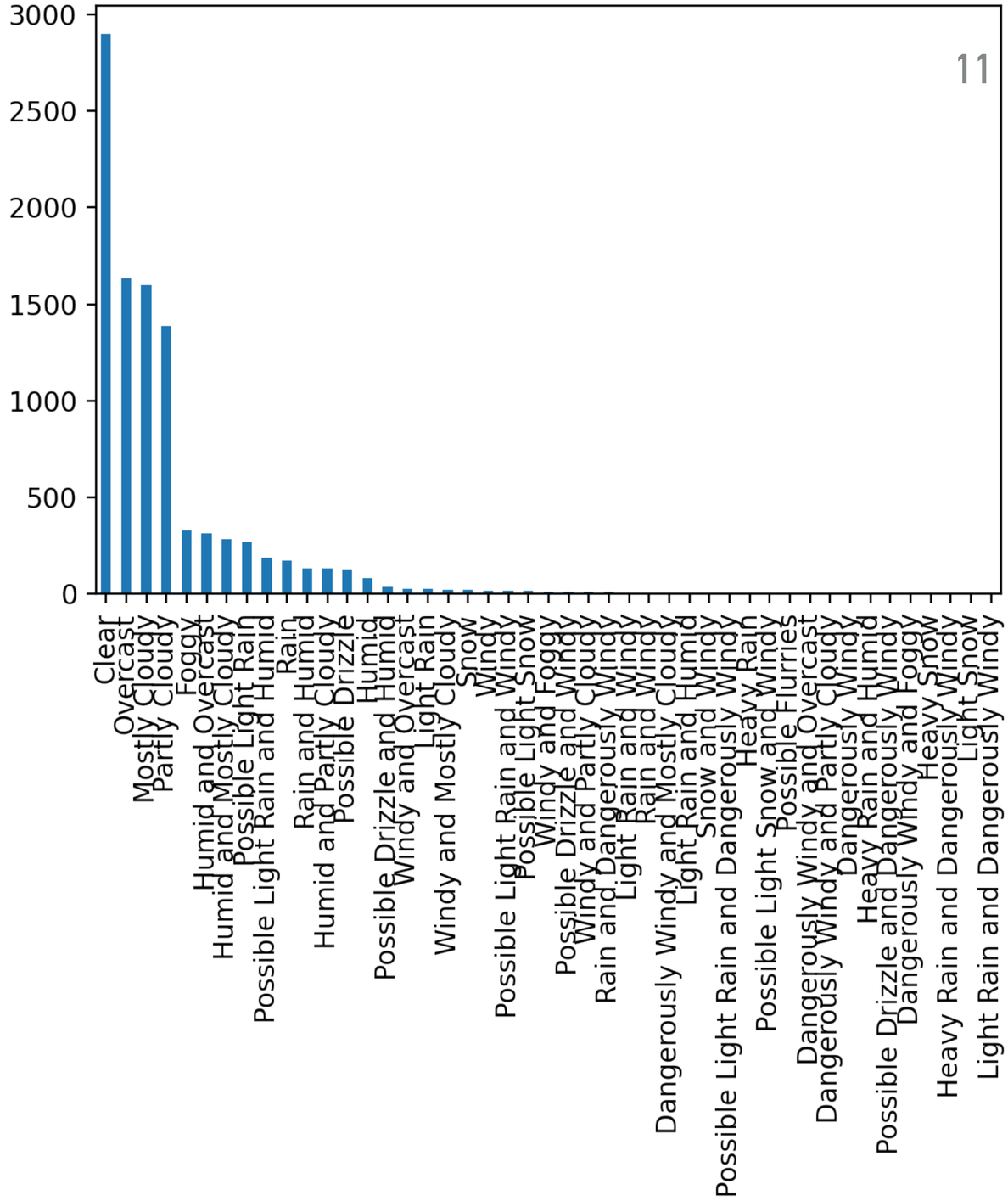
CÁC THÔNG SỐ

- Kích thước: 9828 dòng (mẫu) và 19 cột (đặc trưng).
- Số dữ liệu bị lặp: 0
- Số lượng giá trị bị thiếu: 0

TRỰC QUAN HOÁ CỘT SUMMARY

Nhận xét:

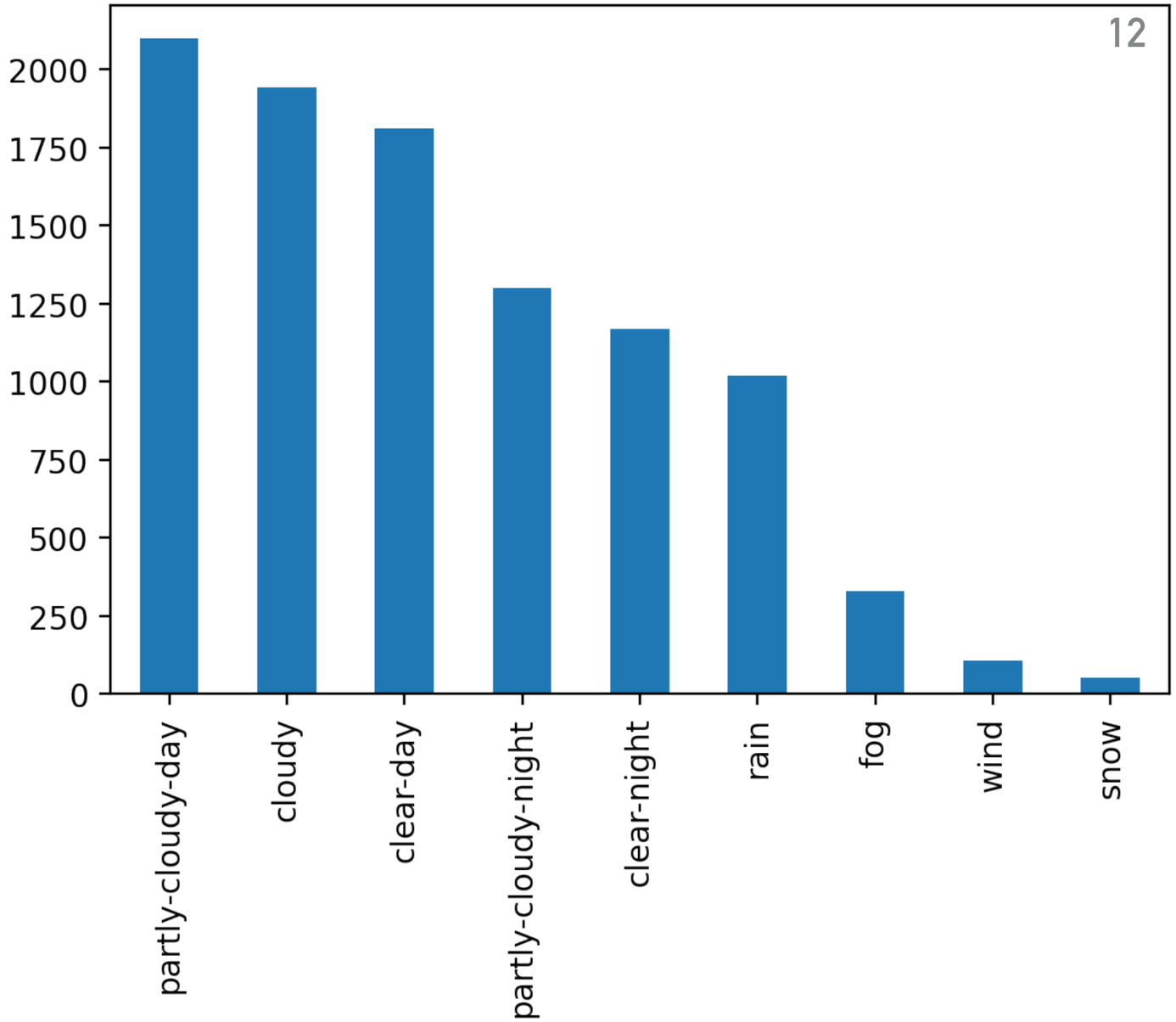
- Dữ liệu bị mất cân bằng.
- Dữ liệu có rất nhiều phân lớp.
- Các phân lớp bị trùng lắp.



TRỰC QUAN HOÁ CỘT ICON

Nhận xét:

- Dữ liệu bị mất cân bằng.
- Dữ liệu có ít phân lớp hơn.
- Các phân lớp khác nhau.



ĐẶT CÂU HỎI

- Dựa vào các số liệu thời tiết hằng ngày như nhiệt độ, độ ẩm, tốc độ gió,... để dự đoán tình trạng thời tiết có thể xảy ra trong vài giờ tới, với:
  - Dự liệu đầu vào: Là các số liệu thời tiết như nhiệt độ, độ ẩm, tốc độ gió,...
  - Dữ liệu đầu ra hay là nhãn sẽ dựa trên cột "icon".
- Mô hình này sẽ cố gắng tập trung dự đoán chính xác các yếu tố thời tiết bất thường như mưa, tuyết, gió, sương để giúp cho người dùng có được sự chuẩn bị thích hợp trong các dạng thời tiết này.

ĐẶT VẤN ĐỀ

“IMBALANCE DATA FOR MULTI CLASSIFICATION.”

Thai Hoang Huy

CÁC PHƯƠNG PHÁP KHẮC PHỤC

- Tăng mẫu huấn luyện.
- Tái tạo mẫu.
- Sự nhạy cảm về chi phí (Cost Sensitivity).

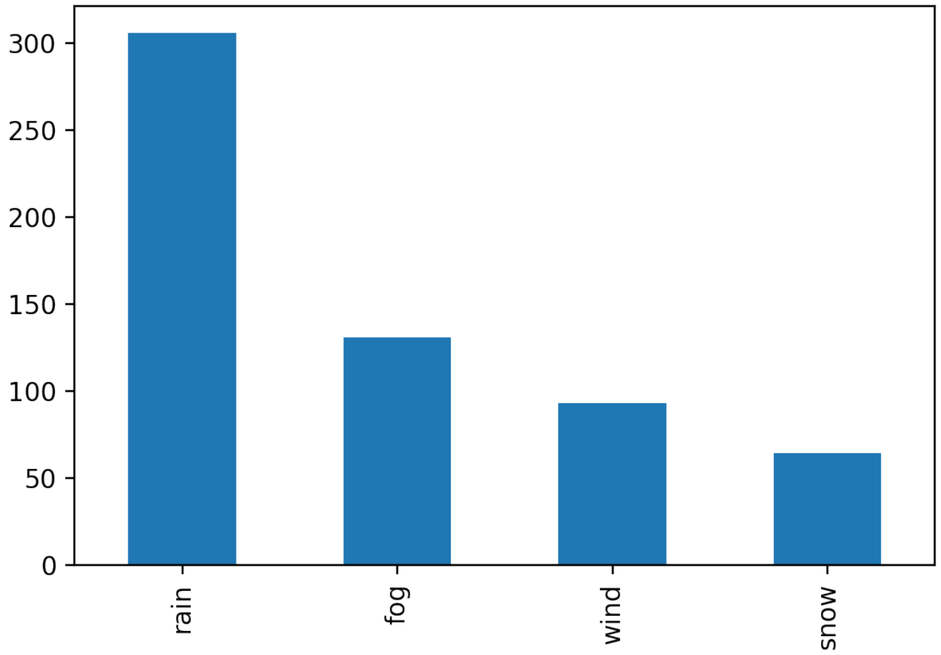
BỔ SUNG DỮ LIỆU

PHƯƠNG PHÁP

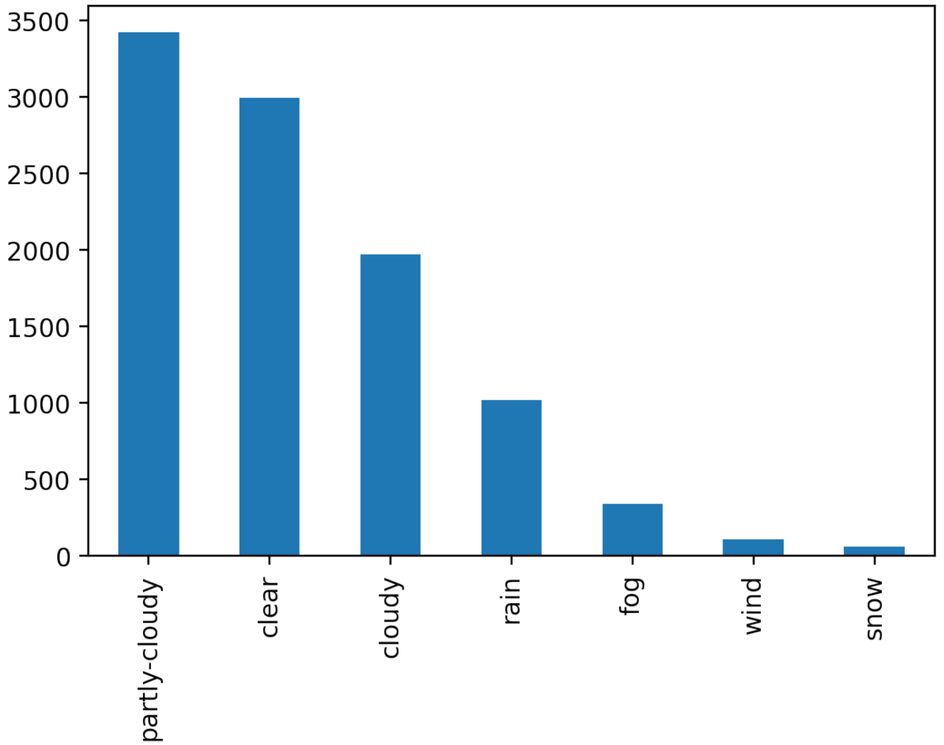
- Việc thu thập dữ liệu thiểu số dựa trên vị trí và thời điểm.
- Loại bỏ các dữ liệu thuộc các lớp partly-cloudy-day, partly-cloudy-night, cloudy, clear-day và clear-night.

CHÚNG TA CÓ GÌ ?

DỮ LIỆU BỔ SUNG



DỮ LIỆU SAU KHI ĐƯỢC GỘP



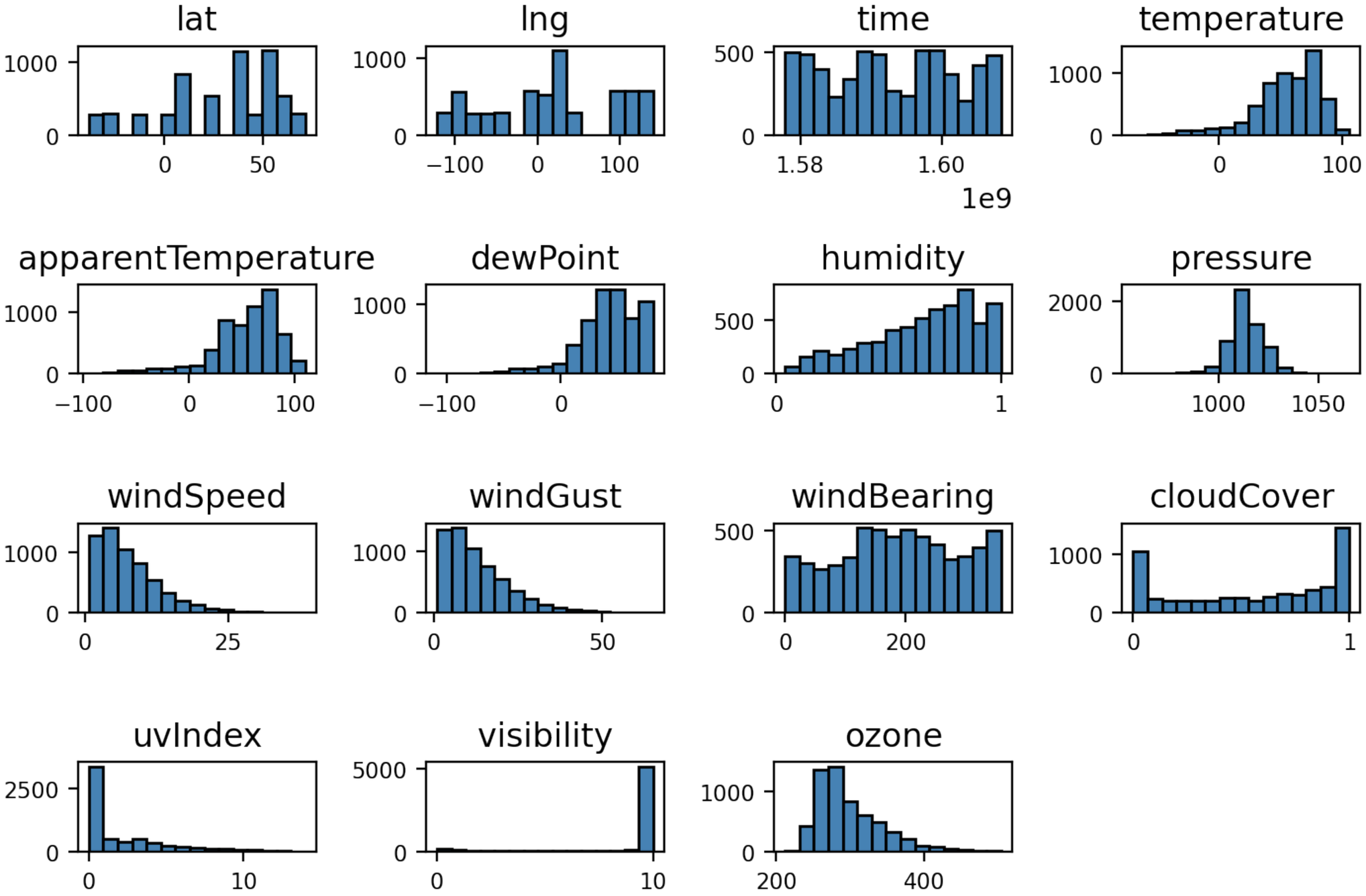
NHẬN XÉT

- Kết quả thu được là không cao (cụ thể chỉ được thêm 594 mẫu thiểu số).
- Việc thu thập những liệu bất thường này diễn ra trong nhiều ngày và tất cả các key API đều quá lượt truy cập.
- Ít cải thiện độ cân bằng của tập dữ liệu.

TIỀN XỬ LÝ
&
KHÁM PHÁ DỮ LIỆU TRÊN TẬP HUẤN LUYỆN

CÁC BƯỚC CƠ BẢN

- Chuẩn hoá cột dữ liệu phân lớp (icon).
- Tách tập dữ liệu theo tỉ lệ 6:2:2 với tham số stratify.
- Chuẩn hoá dữ liệu dạng số bằng Mean Centering và Standardizing.


lat
1000
0
0
50
lng
1000
0
−100
0
100
time
500
0
1.58
1.60
1e9
temperature
1000
0
0
100
apparentTemperature
1000
0
−100
0
100
dewPoint
1000
0
−100
0
humidity
500
0
0
1
pressure
2000
0
1000
1050
windSpeed
1000
0
0
25
windGust
1000
0
0
50
windBearing
500
0
0
200
cloudCover
1000
0
0
1
uvIndex
2500
0
0
10
visibility
5000
0
0
10
ozone
1000
0
200
400

ĐỘ TƯƠNG QUAN GIỮA CÁC THUỘC TÍNH

- Được dùng để đo độ phục thuộc của hai thuộc tính.
- $r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$
  - n là số lượng mẫu trong tập dữ liệu.
  - σ_A, σ_B tương ứng với độ lệch chuẩn của A, B.
  - a_i, b_itương ứng với các giá trị trong thuộc tính A, B.
  - $\bar{A}, \bar{B}$ tương ứng với giá trị trung bình của A, B.

BIỆN LUẬN ĐỘ TƯƠNG QUAN [6]

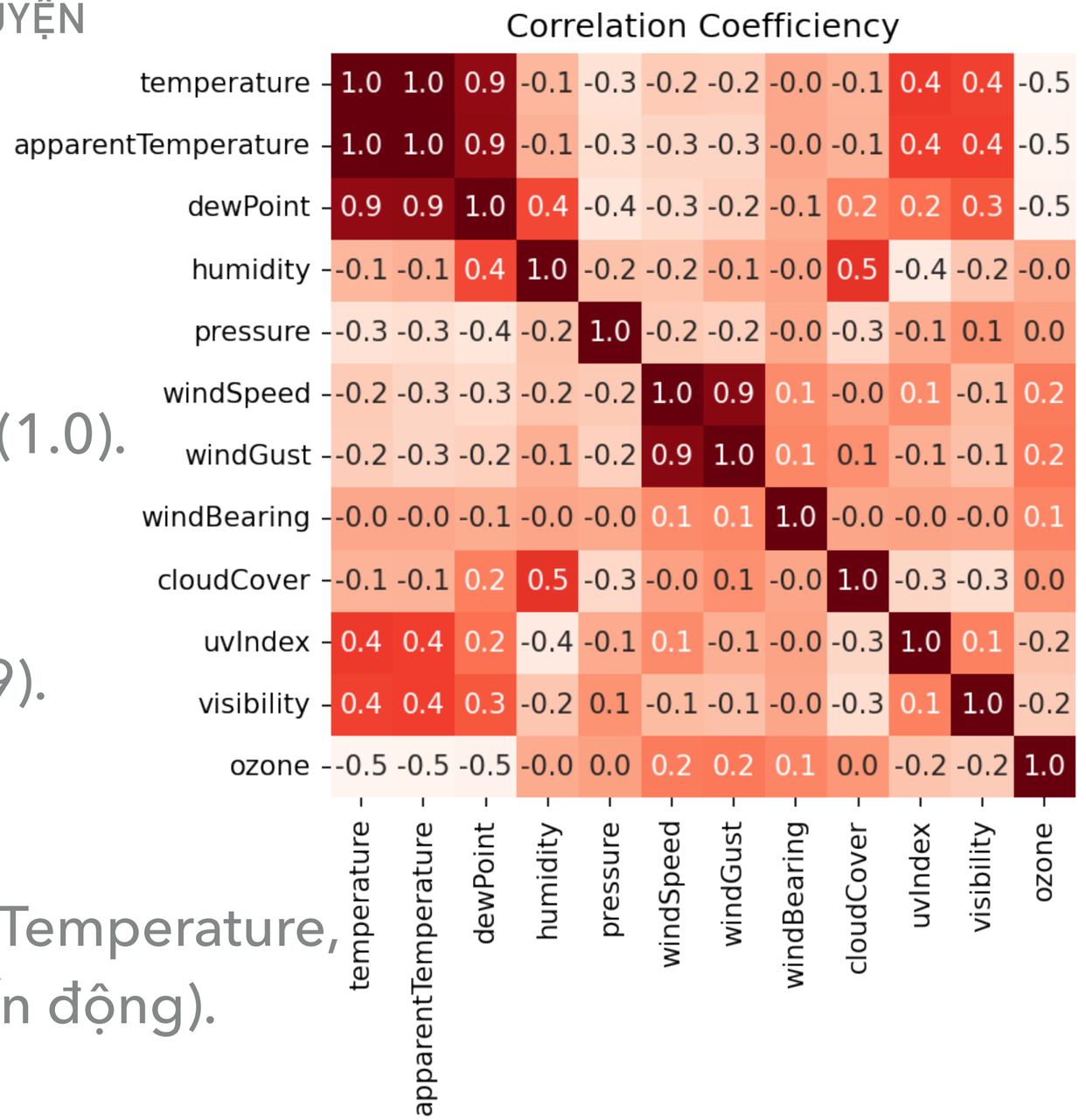
- $r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$
- $r_{A,B}$ có giá trị nằm trong khoảng $[-1,1]$:
  - $r_{A,B} > 0$, thuộc tính A, B có quan hệ dương với nhau tức là nếu A tăng thì B tăng. Giá trị $r_{A,B}$ càng lớn thì mức độ phụ thuộc càng cao, suy ra việc loại bỏ A hoặc B sẽ giúp giảm chiều dữ liệu.
  - $r_{A,B} = 0$, thuộc tính A, B không liên quan đến nhau.
  - $r_{A,B} < 0$, thuộc tính A, B có quan hệ âm với nhau tức là nếu A tăng thì B giảm và ngược .

KẾT QUẢ

Các cặp thuộc tính có quan hệ dương cao:

- temperature và apparentTemperature (1.0).
- temperature và dewPoint (0.9).
- apparentTemperature và dewPoint (0.9).
- windSpeed và windGust (0.9).

- Loại bỏ các thuộc tính như sau: apparentTemperature, dewPoint và windGust (do mang tính biến động).



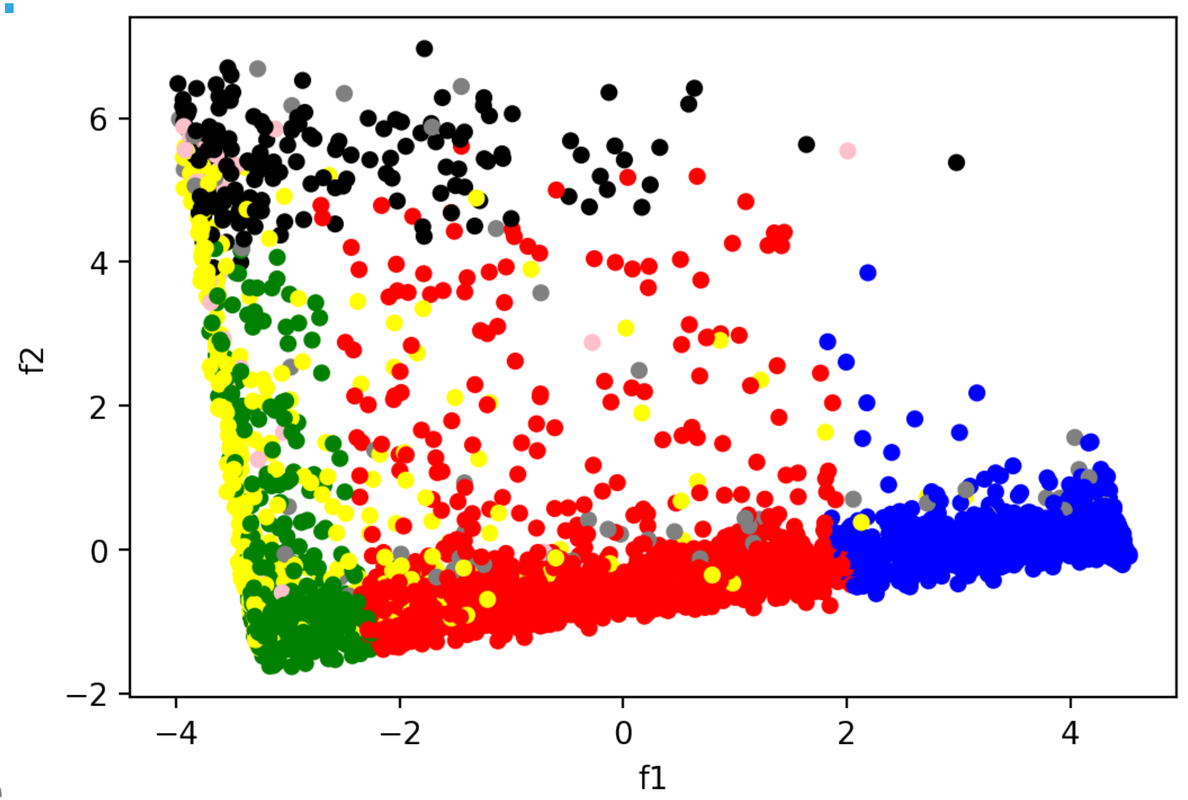
LINEAR DISCRIMINANT ANALYSIS

- Thuật toán rút trích đặc trưng ẩn của tập dữ liệu với mục tiêu: "Cùng gần, khác xa".
- Phương pháp LDA gồm 2 giai đoạn:
  - Giai đoạn 01: Rút gọn số chiều của dữ liệu đầu vào.
    - 1. Tính ma trận phân tán giữa các nhóm (dựa trên kỳ vọng của từng nhóm):
      $S_b = \sum_{i=1}^{C} (\mu_i - \mu)(\mu_i - \mu)^T = H_b H_b^T.$
    - 2. Tính ma trận phân tác tích luỹ tương ứng với từng nhóm (dựa vào mẫu huấn luyện):
      $S_w = \sum_{i=1}^{C} \sum_{j=1}^{M_i} (x_j - \mu_j)(x_j - \mu_j)^T = H_w H_w^T.$
    - 3. Xây dựng hàm tiêu chí tách lớp ($max(H_b), min(H_w)$): $w = argmax(\frac{trace(G^T S_b G)}{trace(G^T S_w G)})$.
  - Giai đoạn 02: Dự đoán nhãn của mẫu dữ liệu nhập (so sánh với vector trung bình của từng nhóm - gần vector trung bình nhất).

PHÂN BỐ CỦA DỮ LIỆU TRÊN TẬP HUẤN LUYỆN

Nhận xét:

- Có sự phân biệt ở các lớp đỏ, xanh dương.
- Ở các lớp còn lại vẫn còn chồng lấn các điểm dữ liệu với nhau.
- Tăng số lượng chiều lên có thể dữ liệu sẽ phân biệt tốt hơn.



TÁI TẠO MẪU [4]

- Undersampling: xoá mẫu phổ biến.
  - Tính toán K mẫu phổ biến và từ K mẫu đó chọn ngẫu nhiên mẫu để xoá.
  - Vấn đề: Xoá mất đi dữ liệu quan trọng và giảm tỉ lệ biểu diễn.
- Oversampling: tăng mẫu thiểu số.
  - Chọn bất kỳ mẫu dữ và thêm vào tập dữ liệu cho đến khi nào đạt trạng thái cân bằng.
  - Vấn đề: có thể dẫn đến tình trạng quá khớp dữ liệu.

ADASYN [4]

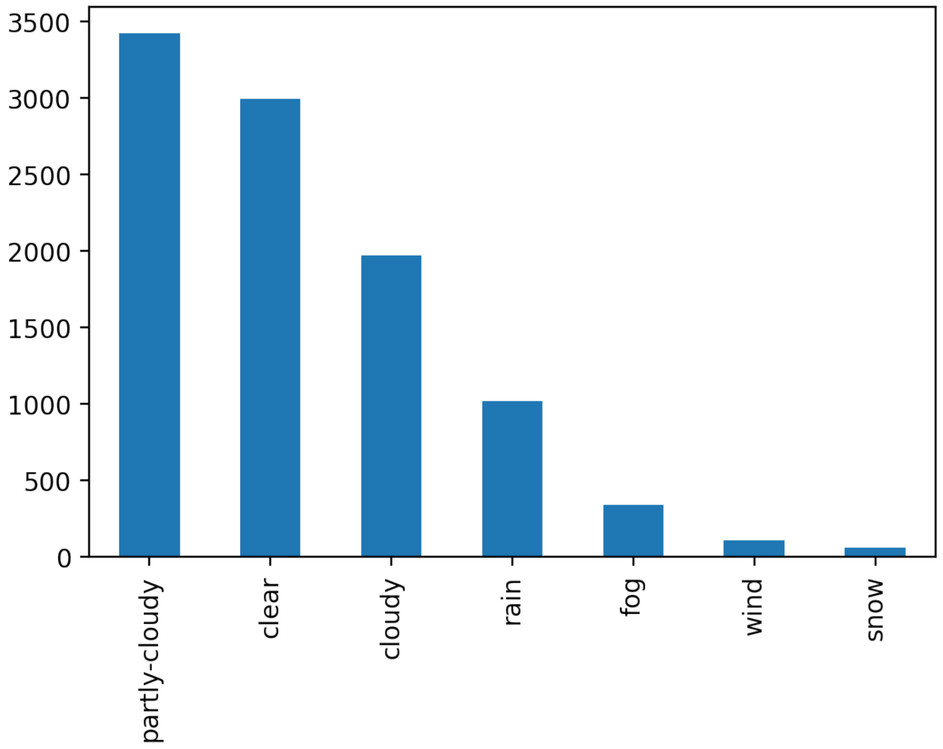
- Dựa trên thuật toán tổng hợp mẫu nhân tạo và cho phép tái tạo thích nghi số lượng mẫu khác nhau dữ trên phân bố của chúng.
- Dữ liệu đầu vào: m mẫu với các giá trị x_i, y_i với x_i là các thuộc tính của mẫu và y_i là nhãn tương ứng với từng mẫu.
- Kết quả đầu ra: tập dữ liệu sau khi được cân bằng.

THUẬT TOÁN [4], [5]

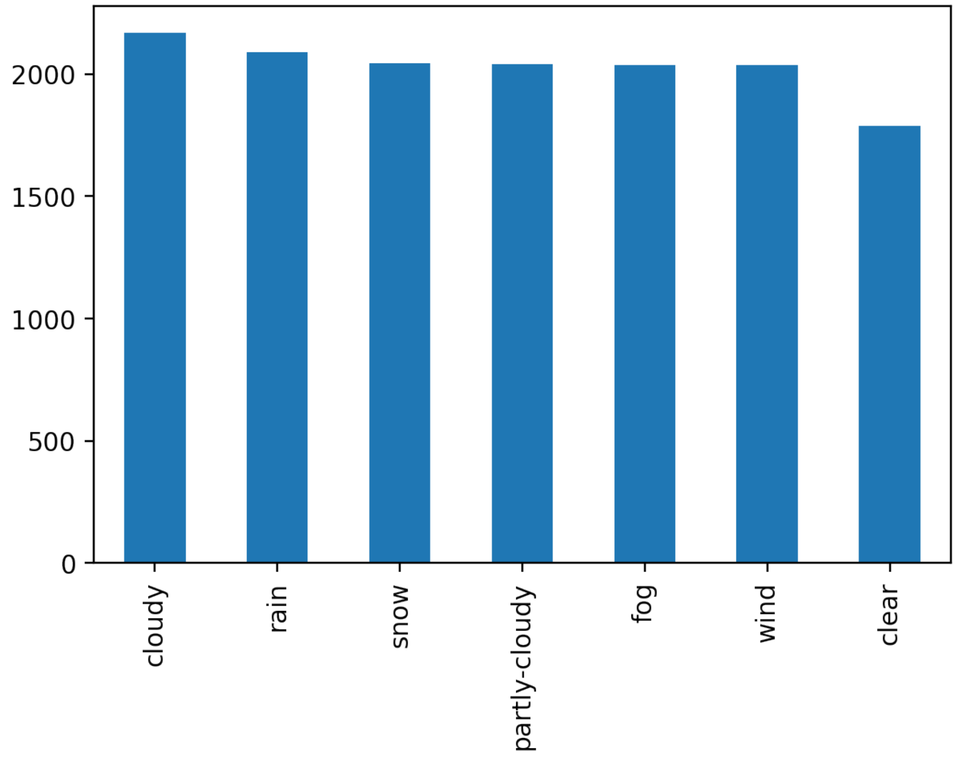
- 1. Tính toán tỉ lệ phân lớp $d = \frac{m_r}{m_x}$ với m_r là số lượng mẫu cực tiểu và m_x là số lượng mẫu cực đại.
- 2. Nếu $d < d_x$ (d_x là siêu tham số dùng để giới hạn tỉ lệ cân bằng mẫu) thì thực hiện tiếp bước 3 và ngược lại thì kết thúc thuật toán.
- 3. Tìm số lượng mẫu cần tái tạo: $G = (m_x - m_r) \times \beta$ với β là siêu tham số cân bằng ($\beta = 1$, cân bằng hoàn toàn giữa tập thiểu số và tập phổ biến).
- 4. Tìm K láng giềng gần nhất của từng mẫu trong tập thiểu số bằng khoảng cách Euclide (các láng giềng này có thể không thuộc tập thiểu số) và tính $r = \frac{\Delta_i}{K}$ với Δ_i là số phần tử láng giềng không thuộc lớp thiểu số.
- 5. Chuẩn hoá $r_x = \frac{r_i}{\sum_{j=1}^{K} r_j}$ thành phân bố mật độ.
- 6. Tính toán số mẫu nhân tạo cho từng mẫu trong tập thiểu số: $g_i = r_x \times G$.
- 7. Tái tạo g_i mẫu cho từng mẫu x_i trong tập thiểu số bằng cách chọn ngẫu nhiên từ các mẫu láng giềng (x_{zi}) của x_i (thuộc lớp thiểu số). Mẫu mới được tạo ra theo công thức: $s_i = x_i + (x_{zi}) + x_i) \times \alpha$ với α có giá trị từ 0 đến 1.

CHÚNG TA CÓ GÌ ?

DỮ LIỆU BAN ĐẦU



DỮ LIỆU SAU KHI TÁI TẠO



CHUẨN HOÁ DỮ LIỆU ?

	temperature	humidity	pressure	windSpeed	windBearing	cloudCover	uvIndex	visibility	ozone
237	1.049574	-1.329572	-1.256298	0.333737	-0.712215	-0.806406	1.550217	0.340715	-0.199109
322	-0.430524	-0.950154	0.432492	-0.106249	0.757232	-0.752026	-0.660558	0.340715	1.021934
135	-0.239348	-0.317791	0.410416	0.512007	0.517322	-0.806406	0.076367	0.340715	0.618781
10	-1.005949	0.778306	0.112395	-0.379344	1.337014	-0.887976	-0.660558	-3.572616	2.845388
37	-0.865981	0.693991	0.874006	0.015127	-0.282377	-0.588885	-0.660558	-0.532862	0.880598
...	...	...	...	...	...	...	...	...	...
376	1.049194	-1.793305	-0.483649	0.965268	-1.591884	0.009297	1.181755	0.340715	-0.699575
296	0.744223	1.157724	-0.461573	-1.147422	0.747236	0.525909	-0.660558	0.340715	-0.810789
345	0.282214	1.368512	0.068243	-0.904671	-1.301993	1.096901	-0.292095	0.340715	-0.025337
392	-0.276521	0.736149	-3.221033	2.258675	0.167454	1.178471	-0.660558	-2.423879	-0.226913
414	-0.013654	-1.160942	0.465606	-1.024150	-1.441941	-1.540539	-0.292095	0.340715	-0.588360

5896 rows × 9 columns

XÂY DỰNG MÔ HÌNH

MÔ HÌNH MULTI LAYER PERCEPTRON

- Mô hình Multi Layer Perceptron là một chuỗi các mô hình perceptron kết hợp lại với nhau để dự đoán kết quả đầu ra Y dựa vào dữ liệu đầu vào X.
- Các bước trong MLP:
  - Mô hình MLP gồm L lớp ẩn.
  - Hàm kích hoạt (activation function).
  - Backpropagation.

MÔ HÌNH MULTI LAYER PERCEPTRON

- Lợi ích của MLP:
  - Có khả nặng học được các biểu diễn phi tuyến.
  - Có khả năng học được mô hình trong thời gian thực.
- Rào cản của MLP:
  - Mô hình không thể hội tụ nếu trong tập dữ liệu có nhiều hơn một cực trị cục bộ và với trọng số khởi tạo khác nhau sẽ cho giá trị khác nhau.
  - Mô hình yêu cầu nhiều giá trị siêu tham số.
  - Nhạy cảm với việc mở rộng các đặc trưng.

THAM SỐ TRONG MLP

- Lớp ẩn: 20 lớp
- Hàm kích hoạt: ReLU (Rectified Linear Unit)
  - Công thức: $f(s) = max(0,s)$.
  - Đơn giản giúp tiết kiệm thời gian trong quá trình backprogation và hạn chế mất mát.
  - Giúp mô hình hội tụ nhanh hơn.

LIMITED-MEMORY BROYDEN FLETCHER GOLDFARB SHANNO [2]

- Thuật toán dựa trên thuật toán Quasi-Newton (trọng số trong mô hình được cập nhật bằng gradient và ma trận Hess).
- Hàm ước lượng ma trận Hess:
  - $\mathbf{H}_{n+1}^{-1} = \left(I - \rho_n y_n s_n^T\right) \mathbf{H}_n^{-1} \left(I - \rho_n s_n y_n^T\right) + \rho_n s_n s_n^T$
  - Trong đó:
    - s_n hiệu kết quả input.
    - y_n là hiệu gradient của phần tử trước đó.
- Mô hình này được dùng nhiều trong các trường hợp có dữ liệu ít vì hội tụ nhanh hơn và cho kết quả tốt hơn.

MÔ HÌNH SUPPORT VECTOR CLASSIFICATION

- Mô hình này thường được dùng cho các bài toán phân tách hai lớp sao cho khoảng cách gần nhau nhất của hai điểm thuộc hai lớp là xa nhau nhất có thể.
- Ưu điểm [3]:
  - Cho kết quả tốt khi có sự rạch ròi rõ ràng giữa các lớp.
  - Hiệu quả cao trong không gian nhiều chiều.
  - Hiệu quả trong trường hợp số chiều nhiều hơn số phân lớp.
  - Mô hình SVM quản lý bộ nhớ hiệu quả.
  - Mô hình này có hỗ trợ cập nhật trọng số w dựa trên số lượng của từng phân lớp (cost sensitivity).
- Nhược điểm [3]:
  - Không thích hợp đối với khối lượng dữ liệu lớn.
  - Dễ bị ảnh hưởng bởi nhiễu hay các lớp bị trùng lắp nhau.
  - Không thể mô hình hoá các tập dữ liệu có số thuộc tính nhiều hơn số lượng mẫu.

CÁC BƯỚC CHẠY CỦA MÔ HÌNH SVC

- 1. Chọn hàm kernel K.
- 2. Chọn giá trị điều khiển biến trượt để tránh tình trạng quá khớp dữ liệu trên tập huấn luyện: C.
- 3. Bài toán tối ưu hoá bậc hai để tìm tham số cho các vector hỗ trợ: $w = \sum_{i \in SV} \alpha_i y_i x_i$.
- 4. Xây dựng hàm tách lớp từ các vector hỗ trợ: $g(x) = \sum_{i \in SV} \alpha_i x_i^T x + b$.

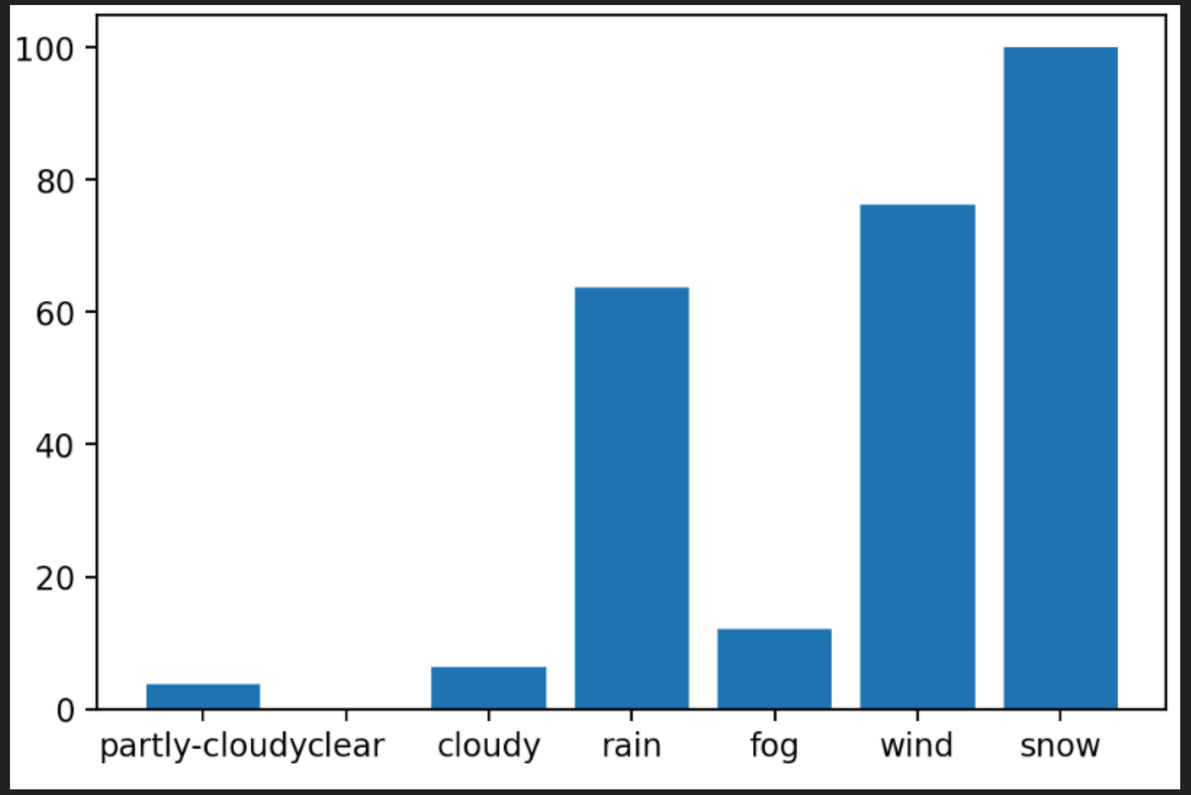
THAM SỐ TRONG SVC

- Hàm kernel:
  - Linear: Hàm tuyến tính với $K(x_i, x_j) = x_i^T x_j$.
  - Poly: Hàm mũ với $K(x_i, x_j) = (1 + x_i^T x_j)^p$.
  - Rbf: Hàm phân phối Guassian với $K(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$.
  - Sigmoid: Hàm phi tuyến với $K(x_i, x_j) = tanh(\beta_0 x_i^T x_j + \beta_1)$.
- Tham số class_weight [1]: "balanced" để cân bằng mức độ quan trọng giữa các lớp khi cập nhật trọng số w trong quá trình học theo công thức: $weight = \frac{nsamples}{nclasses * np.bincount(y)}$.

THỬ NGHIỆM MÔ HÌNH

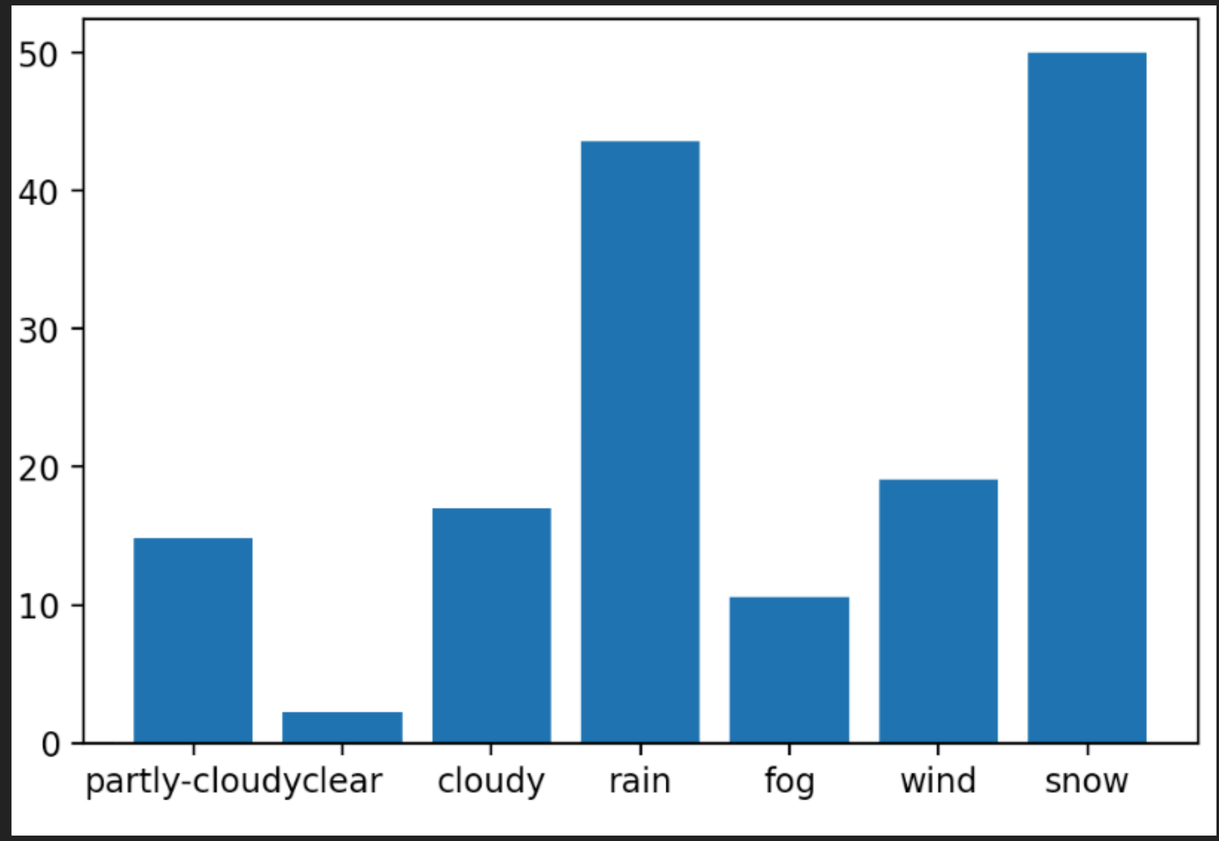
MÔ HÌNH MLP TRÊN DỮ LIỆU BAN ĐẦU

- Độ lỗi tốt nhất: 8.646998982706002.
- Giá trị alpha tốt nhất: 10.
- Độ lỗi trên tập kiểm tra: 10.935910478128179.



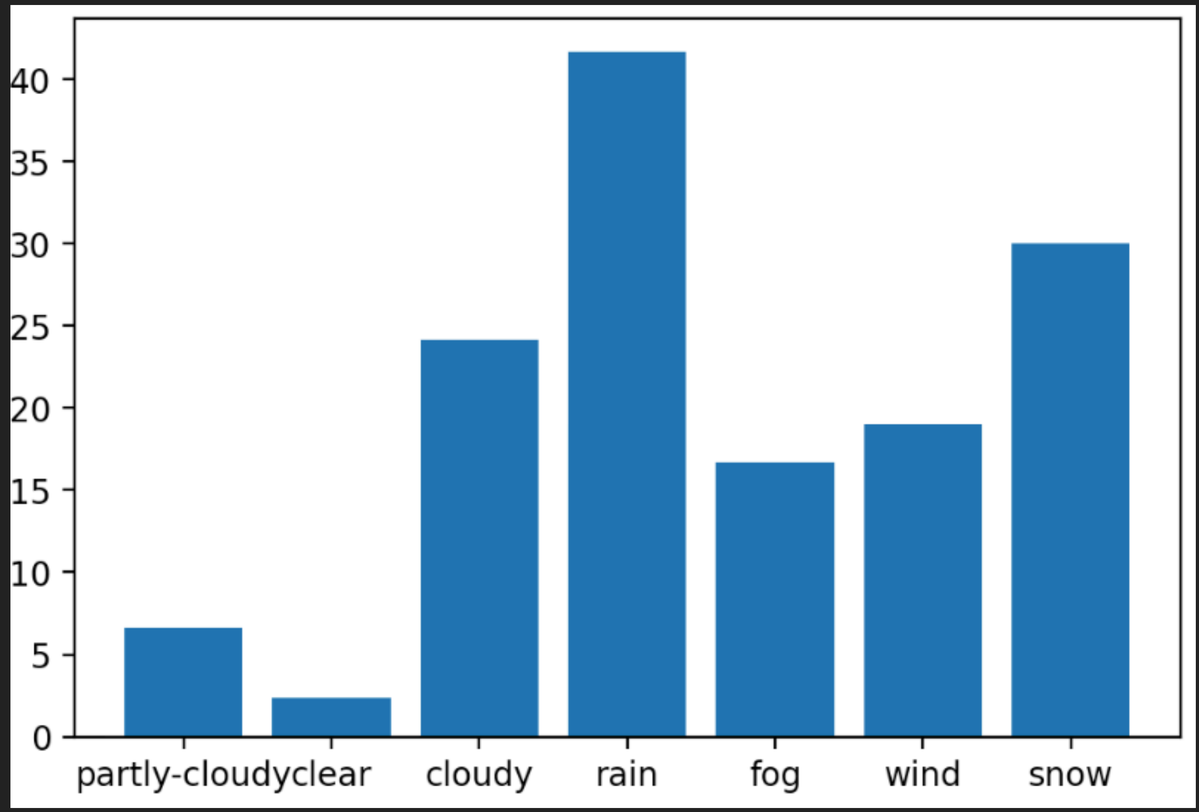
MÔ HÌNH MLP TRÊN DỮ LIỆU TÁI TẠO

- Độ lỗi tốt nhất: 12.71617497456765.
- Giá trị alpha tốt nhất: 0.1.
- Độ lỗi trên tập kiểm tra: 14.496439471007122.



MÔ HÌNH SVC TRÊN DỮ LIỆU BAN ĐẦU

- Độ lỗi tốt nhất: 13.428280773143438.
- Kernel tốt nhất: poly.
- Độ lỗi trên tập kiểm tra: 13.021363173957273.



Ý TƯỞNG CẢI TIẾN

- Thu thập thêm dữ liệu.
- Xây dựng một mô hình phức tạp hơn để xử lý việc mất cân bằng dữ liệu trong đa phân lớp.

TÀI LIỆU THAM KHẢO

- [1]: Book: Imbalanced Classification with Python. Jason Brownlee. Page: 211-221. 2020.
- [2]: Website: aria42. Title: Numerical Optimization: Understanding L-BFGS. 2014. Link:https://aria42.com/blog/2014/12/understanding-lbfgs.
- [3]: Website: Medium. Title: Top 4 advantages and disadvantages of Support Vector Machine or SVM. Author: Dhiraj K. 2019. Link:https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107.
- [4]: Paper: On methods for improving the accuracy of multi-class classification on imbalanced data. Author: Leonid A. Sevastianova, Eugene Yu. Shchetininb. 2020.
- [5]: Website: Medium. Title: Fixing Imbalanced Datasets: An Introduction to ADASYN (with code!). Author: Rui Nian. 2018. Link: https://medium.com/@ruinian/an-introduction-to-adasyn-with-code-1383a5ece7aa.
- [6]: Book: Data Mining Concepts and Techniques. Author: Jiawei Han, Micheline Kamber, Jian Bie. Page: 96. 2012.