ĐỒ ÁN CUỐI KÌ KHOA HỌC DỮ LIỆU

DỰ ĐOÁN ĐỘ ẨM TƯƠNG ĐỐI

Nhóm 11

Nguyễn Văn Hậu	18120359
Nguyễn Tấn Thìn	18120085

Nội Dung

1. Ý tưởng chủ đề
2. Thu thập dữ liệu
3. Khám phá dữ liệu
4. Tách tập, tiền xử lý và khám phá dữ liệu
5. Tiền xử lý và mô hình hóa dữ liệu
6. Nhìn lại quá trình

1. Ý Tưởng Chủ Đề

Câu hỏi: Output - độ ẩm tương đối của ngày mai được tính từ input - các thông số khác (độ ẩm tương đối, nhiệt độ, điểm sương, áp suất, ...) của ngày hiện tại và vài ngày trước như thế nào?

Ý nghĩa: Dự đoán được độ ẩm tương đối, hữu ích trong việc đưa ra các biện pháp phòng chống các hậu quả gây ra do độ ẩm quá cao hoặc quá thấp.

1. Ý Tưởng Chủ Đề

Cảm hứng: Từ thắc mắc về ý nghĩa của độ ẩm tương đối trong thực tiễn, nhóm đã tìm hiểu về độ ẩm tương đối thông qua nhiều nguồn khác nhau trên Internet và mong muốn dự đoán được độ ẩm trong ngày kế tiếp.

2. Thu Thập Dữ Liệu

- Tổng quan về Visual Crossing Weather API
- Cú pháp API
- Cách thu thập dữ liệu

2. Thu Thập Dữ Liệu - Tổng Quan API

- Nguồn API: **Visual Crossing API**

2. Thu Thập Dữ Liệu - Tổng Quan API

- **Visual Crossing API** cung cấp dữ liệu thời tiết trong quá khứ tại một địa điểm giữa hai khoảng thời gian
- Cần có API key để sử dụng
- Mỗi API key chỉ được thu thập **tối đa 1000** kết quả một ngày

2. Thu Thập Dữ Liệu - Cú Pháp API

https://weather.visualcrossing.com/VisualCrossingWebServices/rest/services/timeline/Location/Date1/Date2?key=Your_API_Key

Với

- Location: địa điểm thu thập thông tin
- Date1, Date2: ngày bắt đầu và ngày kết thúc
- Your_API_Key: API key cung cấp khi đăng ký tài khoản

2. Thu Thập Dữ Liệu - Cách Thu Thập

- Thu thập dữ liệu theo từng ngày tại Thành phố Hồ Chí Minh trong khoảng thời gian từ **1/1/2009** đến **31/12/2020**
- Sử dụng nhiều API key để chia ra thu thập dữ liệu theo nhiều lần

3. Khám Phá Dữ Liệu

- Tổng thể toàn bộ dữ liệu ban đầu

	Datetime	Tempmax	Tempmin	Temp	Feelslikemax	Feelslikemin	Feelslike	Dew	Humidity	Precip	Precipcover	Windspeed	Winddir	Windgust	Pressure	Cloudcover	Visibility	Conditions	Icon
0	2009-01-01	31.1	22.1	25.6	34.7	22.1	26.8	22.9	86.50	0.5	4.17	13.0	269.4	NaN	1009.6	69.6	10.7	Rain, Partially cloudy	cloudy
1	2009-01-02	29.7	22.1	25.3	31.4	22.1	26.0	20.5	76.42	0.0	0.00	16.6	107.9	NaN	1010.5	65.0	10.8	Partially cloudy	partly-cloudy-day
2	2009-01-03	28.4	21.1	24.7	30.2	21.1	25.2	20.1	76.12	0.0	0.00	11.2	156.3	NaN	1011.3	78.3	10.7	Overcast	cloudy
3	2009-01-04	29.7	21.1	24.5	32.4	21.1	25.3	20.5	79.16	0.0	0.00	9.4	270.6	NaN	1010.9	66.7	10.6	Partially cloudy	partly-cloudy-day
4	2009-01-05	31.1	21.1	25.9	33.8	21.1	27.0	22.0	80.76	0.0	0.00	14.8	174.4	NaN	1010.1	55.0	7.6	Partially cloudy	partly-cloudy-day

3. Khám Phá Dữ Liệu

- Xử lý thêm cột dữ liệu quá khứ và dữ liệu độ ẩm tương đối ngày kế tiếp

	Datetime	Tempmax	Tempmin	Temp	Feelslikemax	Feelslikemin	Feelslike	Dew	Humidity	Precip	...	TempmaxPrev_7	TempminPrev_7	DewPrev_7	PressurePrev_7
0	2009-01-01	31.1	22.1	25.6	34.7	22.1	26.8	22.9	86.50	0.5	...	NaN	NaN	NaN	NaN
1	2009-01-02	29.7	22.1	25.3	31.4	22.1	26.0	20.5	76.42	0.0	...	NaN	NaN	NaN	NaN
2	2009-01-03	28.4	21.1	24.7	30.2	21.1	25.2	20.1	76.12	0.0	...	NaN	NaN	NaN	NaN
3	2009-01-04	29.7	21.1	24.5	32.4	21.1	25.3	20.5	79.16	0.0	...	NaN	NaN	NaN	NaN
4	2009-01-05	31.1	21.1	25.9	33.8	21.1	27.0	22.0	80.76	0.0	...	NaN	NaN	NaN	NaN

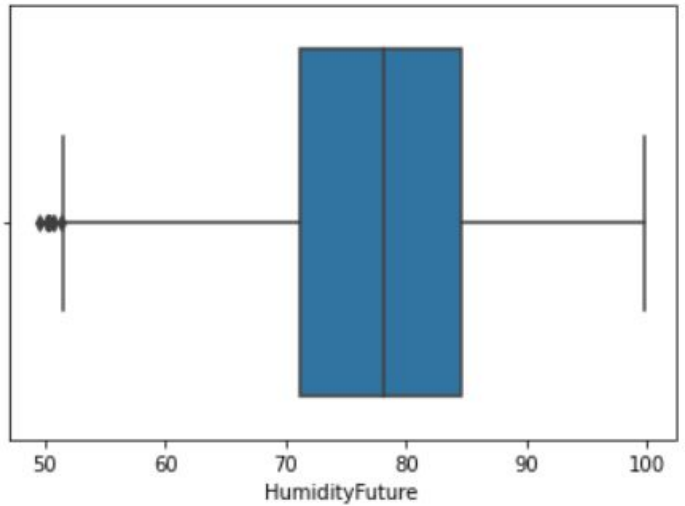
3. Khám Phá Dữ Liệu

- Dữ liệu ban đầu gồm **4383** dòng và **19** cột
- Sau khi thêm cột thì dữ liệu gồm **4383** dòng và **68** cột
- Mỗi dòng là thông tin thời tiết ghi nhận trong một ngày
- Ý nghĩa một số cột

Cột	Ý nghĩa	Đơn vị đo
Datetime	Ngày (yyyy-m-d) ghi nhận thông tin	
Tempmax	Nhiệt độ cao nhất trong ngày	Celsius
Tempmin	Nhiệt độ thấp nhất trong ngày	Celsius
Temp	Nhiệt độ trung bình trong ngày	Celsius
Dew	Điểm sương, là nhiệt độ mà tại đó độ ẩm tương đối của khối không khí đạt 100%	Celsius
Humidity	Độ ẩm tương đối	%
Precip	Lượng mưa trong ngày	mm
Precipcover	Tỉ lệ thời gian mưa trong ngày	%
Windspeed	Tốc độ gió trung bình	kph
Winddir	Hướng gió đo so với hướng Bắc	Degrees
Pressure	Áp suất không khí	millibars
Cloudcover	Tỉ lệ mây bao phủ bầu trời	%
Visibility	Tầm nhìn xa ban ngày	km
Conditions	Hiện tượng thời tiết ghi nhận được như sấm sét, mưa, ...	
Icon	Thời tiết đại diện cho ngày	
HumidityFuture	Độ ẩm tương đối của ngày tiếp theo	%
HumidityPrev_x	Độ ẩm tương đối của x ngày trước	%
TempPrev_x	Nhiệt độ trung bình của x ngày trước	Celsius
TempmaxPrev_x	Nhiệt độ cao nhất của x ngày trước	Celsius
TempminPrev_x	Nhiệt độ thấp nhất của x ngày trước	Celsius
DewPrev_x	Điểm sương của x ngày trước	Celsius
PressurePrev_x	Áp suất không khí của x ngày trước	millibars

4. Tách Tập, Tiền Xử Lý, Khám Phá Dữ Liệu

- Khám phá cột output:
  - Có giá trị số thực
  - Có chứa giá trị rỗng
  - Biểu đồ phân bố như hình
- Dòng mà cột output có giá trị rỗng sẽ bị xóa



4. Tách Tập, Tiền Xử Lý, Khám Phá Dữ Liệu

- Tách tập dữ liệu thành 3 tập:
  - Tập huấn luyện (**70%**) gồm **3066** dòng
  - Tập validation (**15%**) gồm **658** dòng
  - Tập kiểm tra (**15%**) gồm **658** dòng

4. Tách Tập, Tiền Xử Lý, Khám Phá Dữ Liệu

- Kiểu dữ liệu của các cột

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Datetime        3066 non-null   object
 1   Tempmax         3066 non-null   float64
 2   Tempmin         3066 non-null   float64
 3   Temp            3066 non-null   float64
 4   Feelslikemax    3066 non-null   float64
 5   Feelslikemin    3066 non-null   float64
 6   Feelslike       3061 non-null   float64
 7   Dew             3066 non-null   float64
 8   Humidity        3066 non-null   float64
 9   Precip          3066 non-null   float64
 10  Precipcover     3066 non-null   float64
 11  Windspeed       3066 non-null   float64
 12  Winddir         3066 non-null   float64
 13  Windgust        342 non-null    float64
 14  Pressure        2904 non-null   float64
 15  Cloudcover      3066 non-null   float64
 16  Visibility      3066 non-null   float64
 17  Conditions      3066 non-null   object
 18  Icon            3066 non-null   object
 19  HumidityPrev_1  3065 non-null   float64
 20  TempPrev_1      3065 non-null   float64
 21  TempmaxPrev_1   3065 non-null   float64
 22  TempminPrev_1   3065 non-null   float64
 23  DewPrev_1       3065 non-null   float64
 24  PressurePrev_1  2905 non-null   float64
```

4. Tách Tập, Tiền Xử Lý, Khám Phá Dữ Liệu

- Phân bố của các cột dữ liệu dạng số

	Tempmax	Tempmin	Temp	Feelslike	Dew	Humidity	Precip	Precipcover	Windspeed	Winddir	Windgust	Pressure	Cloudcover	Visibility	HumidityPrev_1	TempPrev_1	DewPrev_1	PressurePrev_1
missing_ratio	0.0	0.0	0.0	0.2	0.0	0.00	0.0	0.0	0.0	0.0	88.8	5.3	0.0	0.0	0.00	0.0	0.0	5.3
min	24.0	7.0	22.3	22.4	12.7	49.53	0.0	0.0	7.2	36.2	24.1	1000.6	7.2	5.7	49.53	22.1	12.3	1000.6
lower_quartile	31.9	23.1	26.9	28.5	22.2	71.50	0.0	0.0	15.8	139.3	35.3	1008.1	45.4	9.3	71.30	26.9	22.2	1008.1
median	33.0	24.1	27.9	30.3	23.7	78.30	0.0	0.0	18.4	184.4	40.7	1009.4	54.7	9.9	78.30	27.9	23.7	1009.4
upper_quartile	34.0	25.6	28.8	32.1	24.6	84.50	2.0	4.2	22.3	233.2	48.2	1011.0	66.5	10.3	84.50	28.8	24.6	1011.0
max	38.1	29.9	32.5	39.2	26.9	99.82	151.1	37.5	74.2	343.3	77.8	1017.4	99.6	76.6	99.82	32.5	26.8	1017.4

4. Tách Tập, Tiền Xử Lý, Khám Phá Dữ Liệu

- Phân bố của các cột dữ liệu không phải dạng số

	Datetime	Sunrise	Sunset	Conditions	Icon
missing_ratio	0	0	0	0	0
num_values	1022	768	809	5	5
value_ratios	{'2017-11-10': 0.1, '2017-10-22': 0.1, '2019-12-04': 0.1, '2020-07-11': 0.1, '2017-12-07': 0.1, '2018-12-21': 0.1, '2018-12-13': 0.1, '2018-04-01': 0.1, '2019-08-20': 0.1, '2018-06-30': 0.1, '2020...	{'05:43:32': 1.0, '05:41:50': 0.8, '05:43:33': 0.8, '05:41:51': 0.7, '05:41:52': 0.5, '05:43:30': 0.5, '05:43:34': 0.5, '05:29:39': 0.5, '05:43:06': 0.5, '05:29:42': 0.5, '05:43:24': 0.5, '05:29:4...	{'18:04:04': 3.0, '18:04:03': 1.3, '18:04:05': 1.1, '18:20:03': 0.7, '17:26:37': 0.5, '17:26:38': 0.5, '18:04:06': 0.5, '18:20:04': 0.4, '18:20:00': 0.4, '18:19:58': 0.4, '18:04:02': 0.4, '18:03:5...	{'Partially cloudy': 66.3, 'Rain': 27.9, 'Overcast': 4.7, 'Clear': 1.2}	{'partly-cloudy-day': 57.8, 'rain': 27.1, 'cloudy': 11.9, 'clear-day': 1.7, 'wind': 1.5}

5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

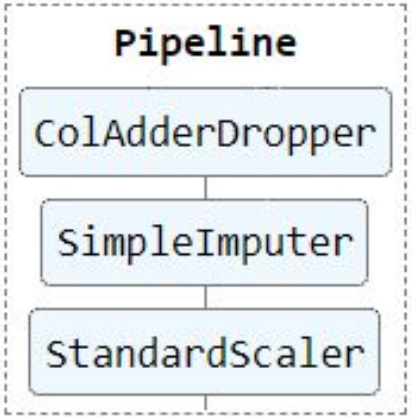
- Class **ColAdderDropper** thêm, xóa một số cột:
  - Tạo thêm Month lấy thông tin tháng từ cột Datetime. Sau đó bỏ cột Datetime.
  - Bỏ các cột Feelslikemax, Feelslikemin, Feelslike, Precip, Precipcover, Winddir, Cloudcover, Visibility, Conditions, Icon vì có lẽ các cột này không mang lại nhiều thông tin hữu ích.
  - Bỏ cột Windgust vì chứa nhiều giá trị rỗng (NaN)

5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Đến đây, dữ liệu hiện chỉ có các cột dạng số (numerical) vì các cột không có dạng số đã bị loại bỏ vì không hữu ích
- Với các cột dữ liệu dạng số ta điền giá trị thiếu bằng **mean** của cột dùng **SimpleImputer**
- Sau khi tất cả các cột đã được điền giá trị thiếu và có dạng số, tiến hành chuẩn hóa tất cả các cột dùng **StandardScaler**

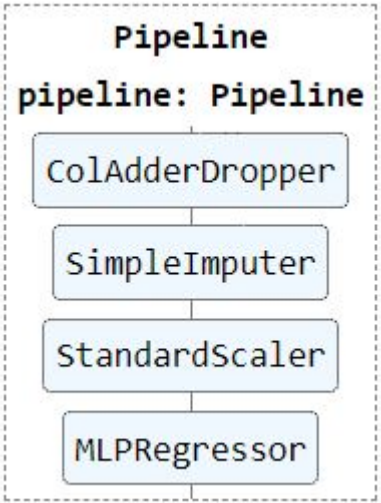
5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Sơ đồ **preprocess_pipeline**



5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Sử dụng **MLPRegressor** với siêu tham số
  - hidden_layer_sizes=(20, 20)
  - activation=relu
  - solver=adam
- Sơ đồ **full_pipeline** như hình

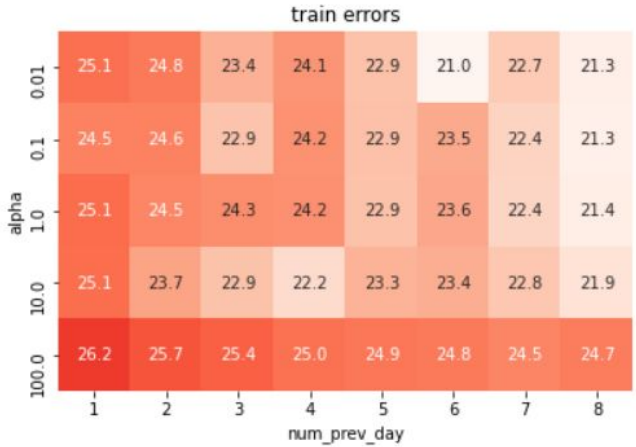


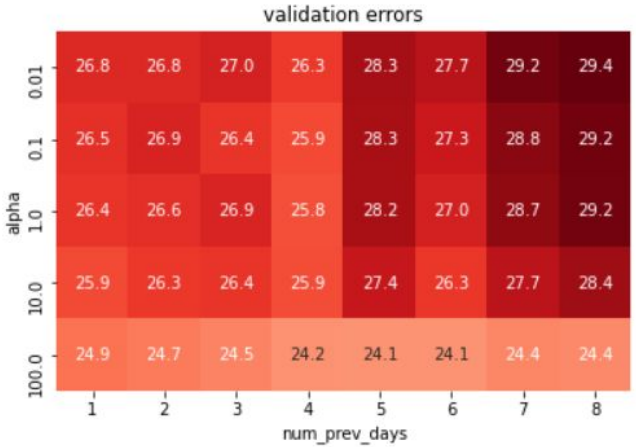
5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Thử nghiệm với nhiều giá trị của **alpha** và **num_prev_day**:
  - **alpha** (mức độ L2 regularization hay weight decay) của MLPRegressor
  - **num_prev_day**: số ngày trong quá khứ mà ta dùng dữ liệu thời tiết của các ngày đó

5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Thử nghiệm với nhiều giá trị của **alpha** và **num_prev_day**





5. Tiền Xử Lý Và Mô Hình Hóa Dữ Liệu

- Bộ siêu tham số tốt nhất là **alpha=100**, **num_prev_day=6**
- Huấn luyện lại mô hình với bộ siêu tham số tốt nhất trên tập dữ liệu kết hợp tập huấn luyện và tập validation
- Kết quả độ lỗi thu được trên tập test là **22.9%**

6. Nhìn lại quá trình làm đồ án

- Về khó khăn bước thu thập dữ liệu:
  - Số lượng API thời tiết khá nhiều nhưng phần lớn là API chỉ cung cấp thời tiết dự báo trong tương lai. Các API cung cấp dữ liệu thời tiết quá khứ thì phần lớn đều tính phí
  - VisualCrossing API khá hạn chế về lượng kết quả được phép thu thập trong ngày

6. Nhìn lại quá trình làm đồ án

- Về khó khăn bước xử lý và mô hình hóa:
  - Khi thêm cột dữ liệu quá khứ vào thì số lượng cột của data khá nhiều, dẫn đến việc độ lỗi trên tập huấn luyện thấp nhưng độ lỗi trên tập validation cao
  - Phải thử nghiệm nhiều solver cũng như kích thước hidden_layer_sizes và các siêu tham số khác nhau để thu được kết quả tốt nhất

6. Nhìn lại quá trình làm đồ án

- Rút ra từ quá trình làm đồ án
  - Quá trình thu thập và xử lý dữ liệu tốn khá nhiều thời gian và công sức
  - Cách xử lý dữ liệu phải phù hợp thì kết quả mới tốt
  - Tìm hiểu về chủ đề đang làm để việc chọn cột input, output thích hợp
  - Không phải model có độ lỗi trên tập validation thấp thì sẽ tốt khi chạy trên tập kiểm tra

6. Nhìn lại quá trình làm đồ án

- Dự định phát triển thêm nếu có thời gian
  - Tăng thêm lượng data, xử lý outliers nếu có
  - Thử nghiệm nhiều mô hình hơn để tìm ra mô hình tốt nhất