ĐỒ ÁN CUỐI KÌ

Chủ đề: Dự đoán một ngày có mưa hay không

Môn học: Nhập môn khoa học dữ liệu
Giảng viên: Trần Trung Kiên

Nhóm 2:

Hà Minh Toàn - 18120599
Phạm viết Xuân - 18120658

1. ĐẶT RA CÂU HỎI

- **Câu hỏi**: Việc trời mưa liệu có thể được đoán từ những thông tin khác hay không?
- **Mục tiêu**: Dùng những kiến thức đã học trong môn học này để dự đoán xem ngày hôm đó có mưa hay không?

=> Khi trả lời được câu hỏi này, người ta có thể biết được trong ngày xác định có mưa hay không để có thể quyết định công việc hoặc những thứ có liên quan.

2. THU THẬP DỮ LIỆU

- Sau một thời gian tìm kiếm api lịch sử thời tiết, hầu hết đều là trả phí. thì nhóm kiếm được 1 api cho free trial premium 60 ngày đầu.
- Historical Weather API Documentation (worldweatheronline.com)

2. THU THẬP DỮ LIỆU

Các tham số trong lệnh của API:

Base URL

Premium API:
HTTP: http://api.worldweatheronline.com/premium/v1/past-weather.ashx
HTTPS: https://api.worldweatheronline.com/premium/v1/past-weather.ashx

Request Parameters

The following parameters may be used:

Parameter	Description	Required	Values
q	Location	Required	See note below.
extra	Include extra information	Optional	See note below.
date	The date to return the weather for.	Required	yyyy-MM-dd (Example: 2009-07-20 for 20 July 2009.)
enddate	If you wish to retrieve weather between two dates, use this parameter to specify the ending date.**Important:** the *enddate* parameter must have the same month and year as the *date* parameter.	Optional	yyyy-MM-dd (Example: 2009-07-22 for 22 July 2009.)
includelocation	Whether to return the nearest weather point for which the weather data is returned for a given postcode, zipcode and lat/lon values.	Optional	Valid values: •yes •no (default)

tp	Specifies the weather forecast time interval in hours. Options are: 1 hour, 3 hourly, 6 hourly, 12 hourly (day/night) or 24 hourly (day average).	Optional	Valid values: •1 •3 (default) •6 •12 •24
format	The output format to return. XML or JSON.	Optional	Valid values: •xml (default) •json
callback	The function name for JSON callback.	Optional	Example: callback=function_name
key	The API key.	Required	Provided when registering your application.

2. THU THẬP DỮ LIỆU

- **Tiến hành lấy dữ liệu từ api:**

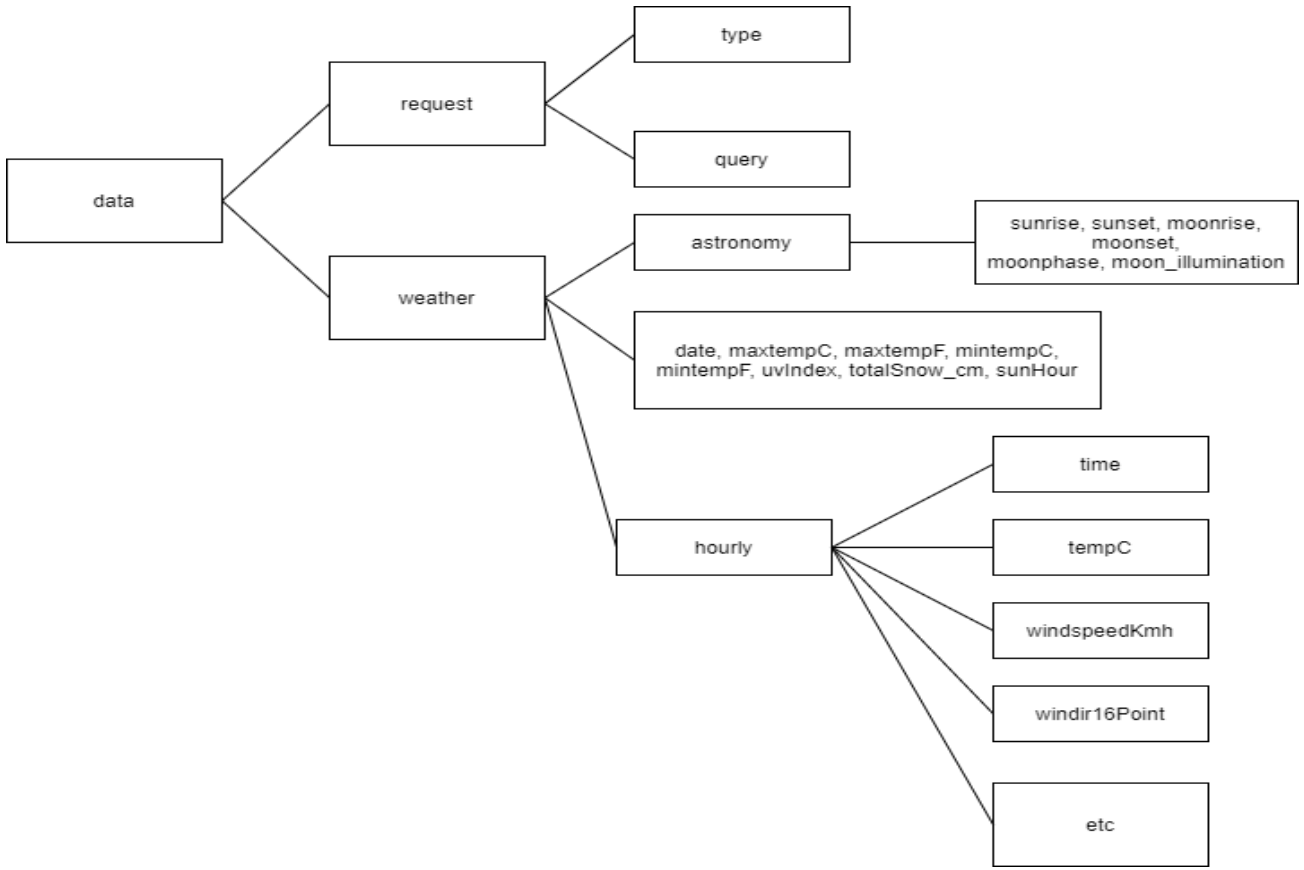
- Mẫu api để lấy dữ liệu từ ngày 1 tới 28 của một tháng ở năm 2013.

- Api yêu cầu date và enddate phải có cùng tháng cùng năm nên muốn lấy dữ liệu của 1 năm thì phải dung for loop.

- http://api.worldweatheronline.com/premium/v1/past-weather.ashx?key=d1f163d12c3742b6a074235821090 1&q=Ho Chi Minh&format=json&date=2013-{tháng}-01&enddate=2013- {tháng}-28&includelocation=yes.

2. THU THẬP DỮ LIỆU

- Cấu trúc của file json api trả về:



2. THU THẬP DỮ LIỆU

- Lựa chọn một số thuộc tính dựa trên ý nghĩa để đưa vào bảng dữ liệu:

```
các cột được chọn:
Ngày + giờ
TempC : Nhiệt độ
Windspeed(km/h): tốc độ gió
Winddir16Point: hướng gió, trên la bàn 16 hướng
WeatherCode: thông tin thời tiết
Humidity
Visibility(km): tầm nhìn
Pressure: áp suất
Cloudcover: tỉ lệ che của mây
```

3. TIỀN XỬ LÝ DỮ LIỆU

- Từ những Weathercode có mưa trong api, gán dữ liệu ngày mưa cho từng dòng dữ liệu.

```
#Những mã weather code của khi có mưa
rain_code = ['389','386','359','356','353','314','311','308','305','302','299','296','293','176']
data_df.loc[~data_df["WeatherCode"].isin(rain_code),"WeatherCode"] = "0"
data_df.loc[data_df["WeatherCode"].isin(rain_code),"WeatherCode"] = "1"
data_df = data_df.rename(columns={"WeatherCode": "Rain Or Not"})
```

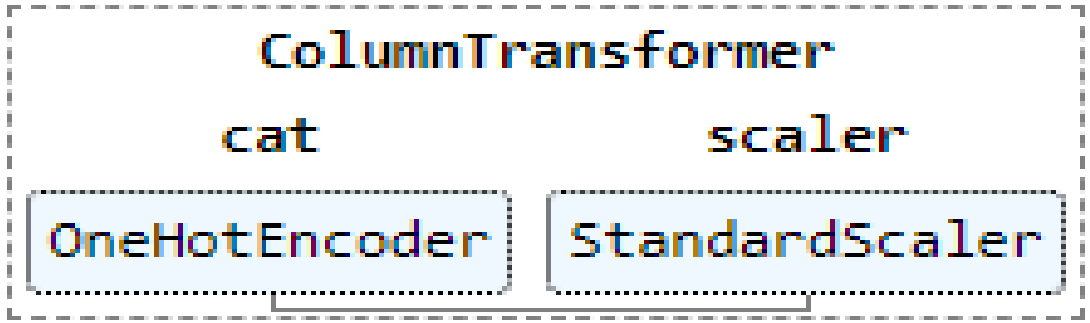
3. TIỀN XỬ LÝ DỮ LIỆU

- **Về mặt dữ liệu**: Bộ dữ liệu thu thập được khá hoàn hảo, không có giá trị thiếu và hầu hết có thể chuyển thành dạng số nên chúng ta không cần thực hiện bước tiền xử lý trên từng cột.
- Cột "**Date + Time**": Loại bỏ
- Cột "**Windir16Point":** Sử dụng OneHotEncoder để mã hóa nó về dạng số.

3. TIỀN XỬ LÝ DỮ LIỆU

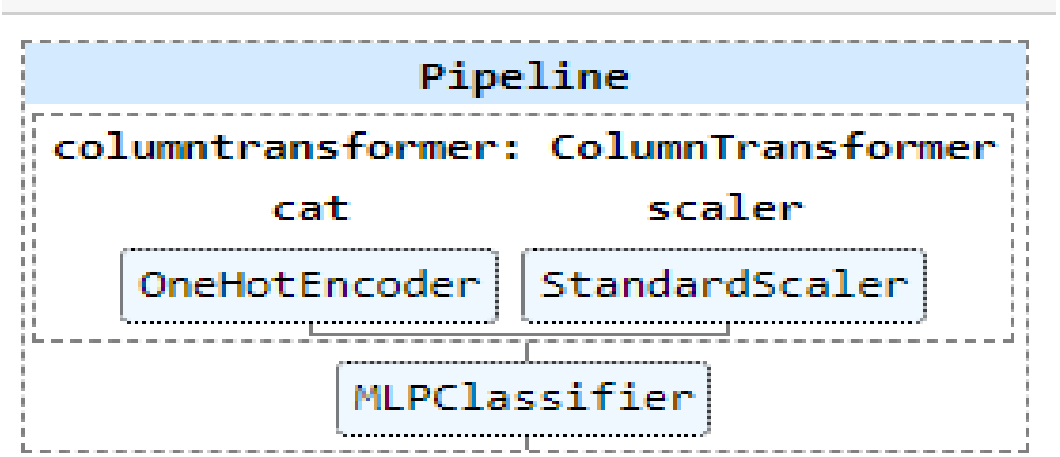
- Tạo một Transformer gồm hai bước: mã hóa cột Windir16Point và StandardScaler các cột còn lại.

```
preprocess_Transformer = ColumnTransformer(
    transformers=[
        ('cat',  OneHotEncoder(handle_unknown='ignore'), ['Windir16Point']),
        ('scaler', StandardScaler(), ['TempC','Windspeed(km/h)','Humidity (%)','Visibility (km)','Pressure (mb)','Cloudcover (%)'])
    ])
preprocess_Transformer
```



4. MÔ HÌNH HÓA DỮ LIỆU

- Mô hình được chọn: Neuron Network.
- Bộ phân lớp được chọn: MLPClassifier.
- Lặp và tính toán độ lỗi với nhiều siêu tham số alpha khác nhau để tìm ra được alpha tốt nhất cho mô hình.
- Fit lại mô hình với alpha tốt nhất vừa tìm được.



5. ĐÁNH GIÁ MÔ HÌNH THU ĐƯỢC

- Kết quả trên tập test cho thấy, độ lỗi đạt 0,721%.
- Với bộ dữ liệu test 416 dòng, kết quả sai khác chỉ có 3.

⇒Kết quả thu được tương đối khả quan.

	TempC	Windspeed(km/h)	Pressure (mb)	Cloudcover (%)	Prediction	Actual
1625	24	11	1007	40	0	1
1263	28	15	1010	26	1	0
1440	31	18	1006	20	0	1

6. KẾT LUẬN

- Từ kết quả của đồ án này, nhóm có thể kết luận được rằng việc mưa trong một ngày có mối quan hệ mật thiết với các thông số thời tiết khác như nhiệt độ, độ ẩm, áp suất không khí, …
- Tuy việc dự đoán này có vẻ thừa vì nếu đã có được những thông số: nhiệt độ, độ ẩm, tốc độ và hướng gió, áp suất không khí,… thì chẳng có lý nào lại không biết được hôm đó có mưa không?
- Nhưng đôi khi vẫn sẽ cần.
- Mục đích chính của việc làm chủ đề này là để củng cố lại kiến thức đã học.

7. Tài liệu tham khảo

- BT03: Các bước làm và đề xuất mô hình của thầy.
- Các nguồn tài liệu trên github, google,…