ĐỒ ÁN CUỐI KÌ

Nhập môn khoa học dữ liệu

Nhóm 19

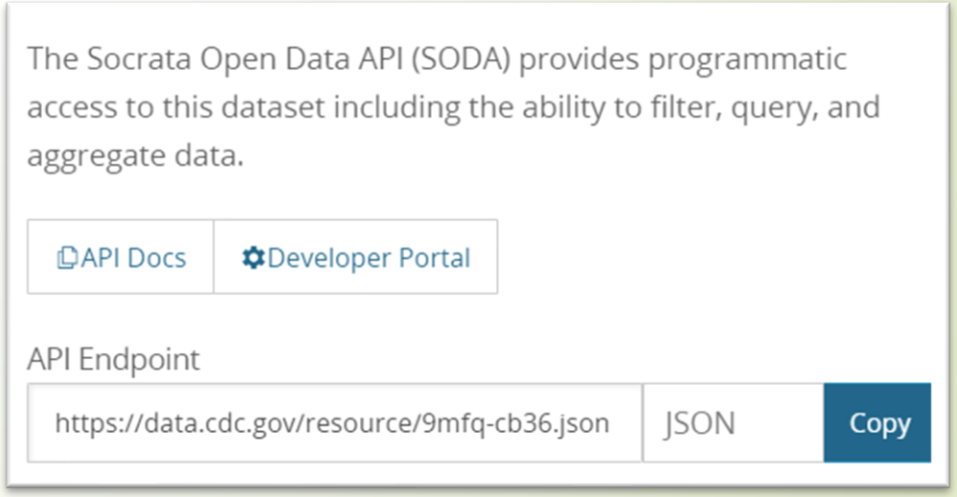
Họ Và Tên	MSSV
Cao Tất Cường	18120296
Hà Văn Duy	18120339

Thu thập dữ liệu thô

Thu thập dữ liệu về số ca mắc và số ca tử vong mỗi ngày của đại dịch COVID-19 trên các bang nước Mỹ (USA).

Dữ liệu được thu thập thông qua API của Trung tâm Kiểm soát Bệnh tật CDC chính thức tại Mỹ.

Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™



API sử dụng để get dữ liệu

Khám phá dữ liệu thô

Nội dung của một vài thuộc tính trong bộ dữ liệu

```
1 data_df.head()
```

	submission_date	state	tot_cases	new_case	pnew_case	tot_death	new_death	pnew_death	created_at	c
0	2020-12-08T00:00:00.000	NM	109947	0.0	0	1756	0.0	0	2020-12-09T14:45:40.234	
1	2021-01-01T00:00:00.000	FL	1300528	0.0	6063	21673	0.0	7	2021-01-02T14:50:51.219	
2	2020-04-30T00:00:00.000	IA	7145	302.0	0	162	14.0	0	2020-05-01T21:00:19.025	
3	2020-06-25T00:00:00.000	NE	18346	125.0	0	260	3.0	0	2020-06-26T19:18:27.809	
4	2020-02-24T00:00:00.000	CA	10	0.0	NaN	0	0.0	NaN	2020-03-26T16:22:39.452	

Một vài thuộc tính chính của dữ liệu thô

Khám phá dữ liệu thô

Các cột của dữ liệu và thông tin mô tả

VARIABLE	DESCRIPTION	TYPE
submission_date	Ngày nộp kết quả	Date & Time
state	Các tiểu bang, vùng lãnh thổ và các khu vực pháp lý của Mỹ	Plain Text
tot_cases	Tổng số lượng ca nhiễm SARS-CoV-2	Number
conf_cases	Tổng số ca nhiễm SARS-CoV-2 đã được xác nhận	Number
prob_cases	Tổng số ca có khả năng cao bị nhiễm SARS-CoV-2	Number
new_case	Số lượng ca nhiễm SARS-CoV-2	Number
pnew_case	Số lượng ca có khả năng bị nhiễm SARS-CoV-2 cao	Number
tot_death	Tổng số ca tử vong do SARS-CoV-2	Number
conf_death	Tông số ca tử vong được xác nhận do SARS-CoV-2	Number
prob_death	Tông số ca tử vong khả năng cao do SARS-CoV-2	Number
new_death	Số lượng ca tử vong mới do SARS-CoV-2	Number
pnew_death	Số lượng ca tử vong mới khả năng cao do SARS-CoV-2	Number
created_at	Ngày tạo dữ liệu	Date & Time
consent_cases	Nếu đồng ý thì confirmed and probable cases có dữ liệu, nếu không thì chỉ total cases có dữ liệu	Plain Text
consent_deaths	Nếu đồng ý thì confirmed and probable deaths có dữ liệu, nếu không thì chỉ total deaths có dữ liệu	Plain Text

Bảng thông tin chi tiết mô tả thuộc tính (Nguồn: CDC)

Đặt câu hỏi

Câu hỏi: Số ca mắc mới trong một ngày bất kì (hợp lệ) là bao nhiêu? Nó phụ thuộc như thế nào vào số ca mắc của những ngày trước đó?

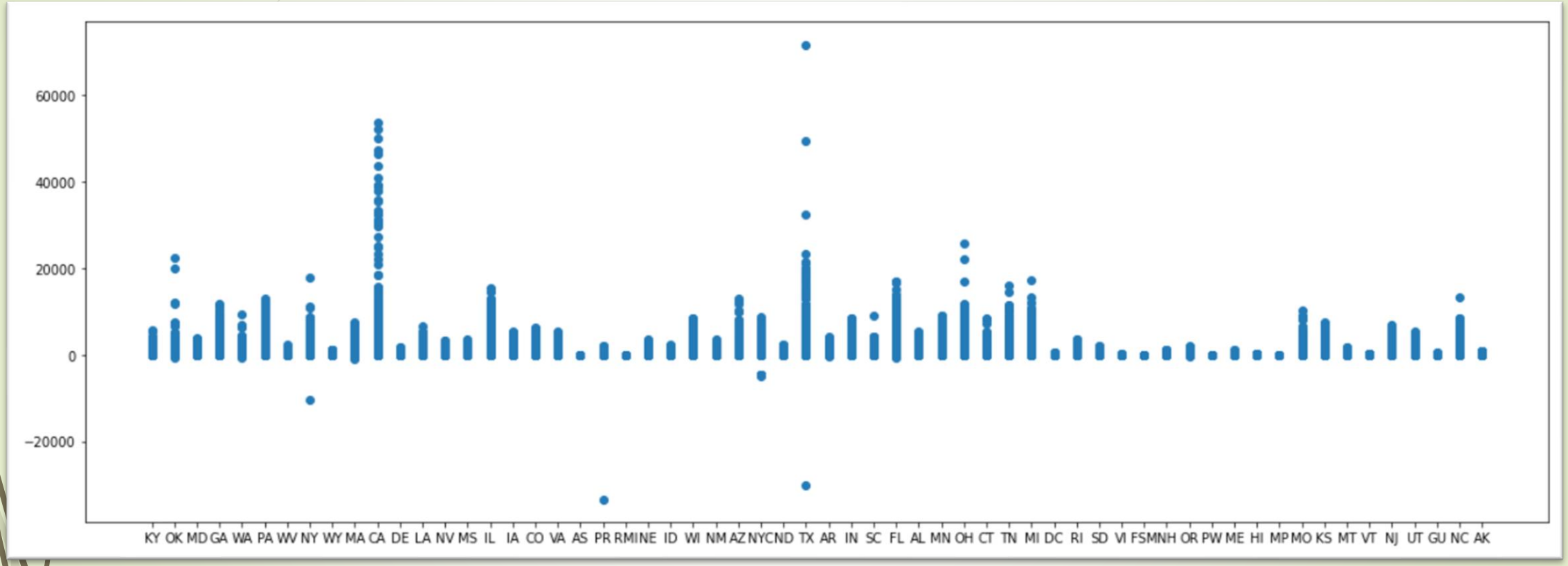
➔ Bài toán Regression

Ý nghĩa:

- ❑ Chỉ ra sự liên hệ giữa số ca mắc hiện tại với số ca những ngày trước đó
- ☑ Thấy khả năng và tần suất lây lan của dịch bệnh
- ❑ Ước lượng các làn sóng dịch tiếp theo trong khu vực
- ❑ Dự đoán tương lai gần (để có biện pháp phù hợp)

Cảm hứng: Bắt nguồn từ thực tế diễn biến đang rất phức tạp của dịch bệnh.

Khám phá cột output (new_case)

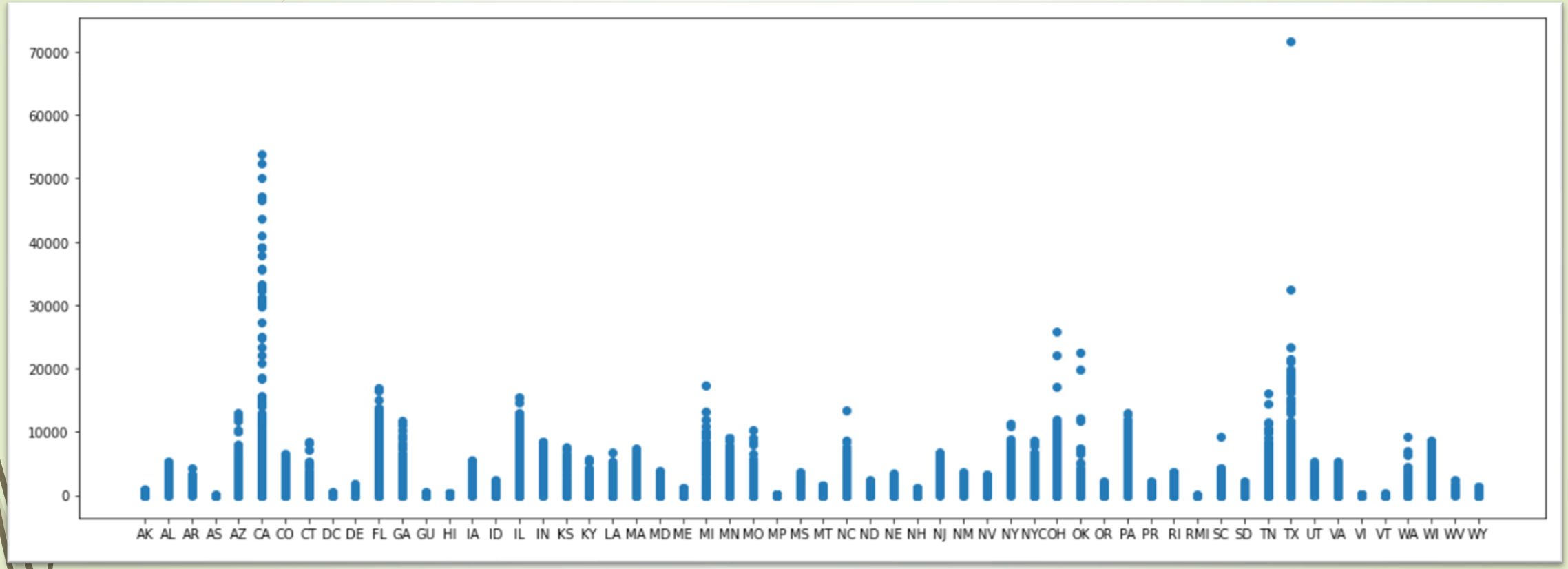


Scatter biểu diễn phân bố của số ca mắc mới (chưa được xử lý) tại mỗi bang

Xử lý dữ liệu thô để phù hợp với bài toán

- Loại bỏ dòng dữ liệu nhiễu
- Thêm các thuộc tính (lịch sử số ca mắc mới của những ngày trước) cho mỗi dòng dữ liệu
- Làm mượt dữ liệu

Xử lý dữ liệu thô để phù hợp với bài toán



Scatter biểu diễn phân bố của số ca mắc mới (sau khi làm mượt) tại mỗi bang

Xử lý dữ liệu thô để phù hợp với bài toán

- Loại bỏ dòng dữ liệu sai (oulier)

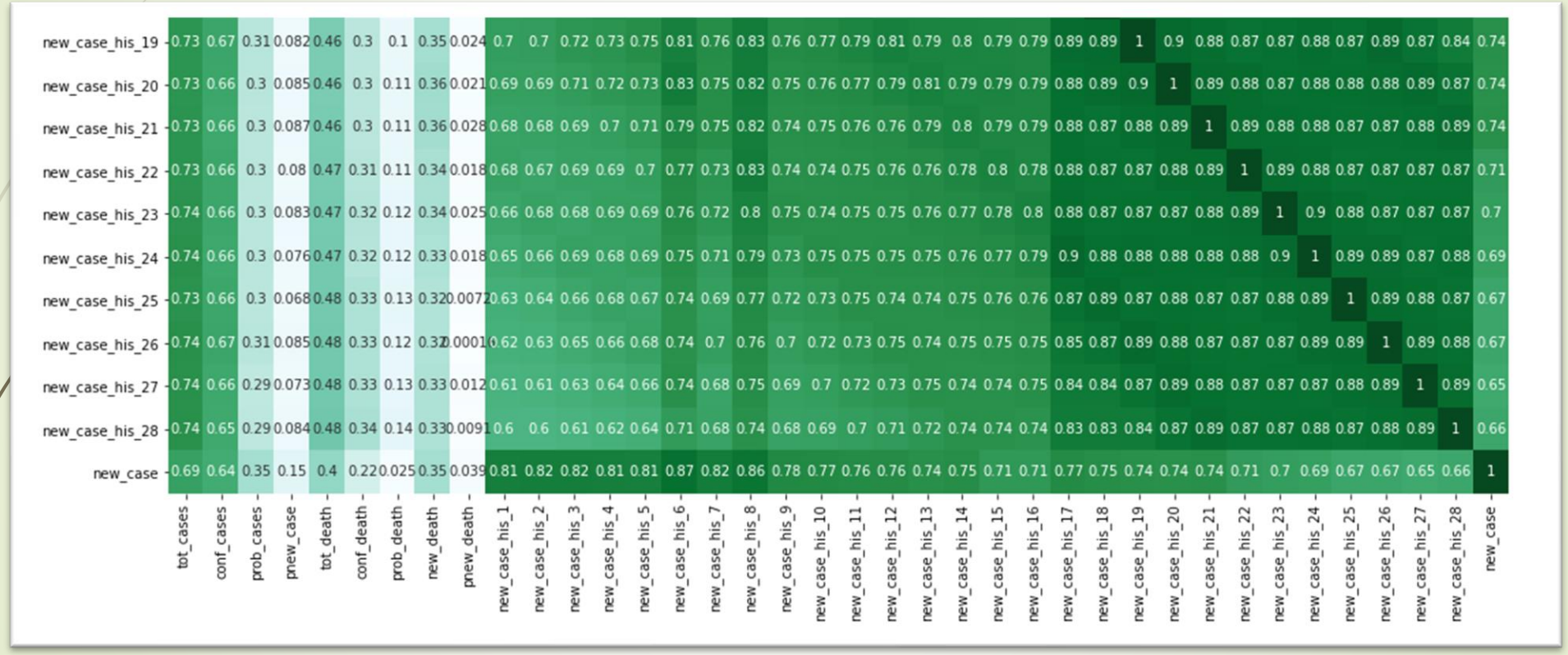
➔ Thu được bộ dữ liệu phù hợp để có thể trả lời câu hỏi.

Tiền xử lý + Mô hình hóa

Tách tập

- Tách dữ liệu dự đoán (*new_case*) ra làm cột riêng là cột output
- Phần còn lại là phần input
- Chia 2 phần trên thành 3 tập: train, validation, test

Xử lý dữ liệu thô để phù hợp với bài toán



Headmap biểu diễn sự tương quan giữa các thuộc tính trong tập dữ liệu

Tiền xử lý + Mô hình hóa

Thêm bớt thuộc tính

Xây dựng Estimator làm các nhiệm vụ sau:

- ❑ Loại bỏ các thuộc tính không có ích tới việc huấn luyện
- ❑ Dựng tham số callback_days
- ❑ Thêm thuộc tính tot_new_case_his

```
ColAdderDropper
ColAdderDropper(callback_days=5)
```

```
[ ]  1 # TEST
     2 estimator.transform(train_X_df).head()
```

	tot_cases	prob_cases	tot_death	new_death	total_new_case_his
15611	80695	3372.0	2399	84.0	3757.0
10698	741	182.0	7	0.0	54.0
3783	1834	NaN	13	6.0	1050.0
4317	220819	4620.0	23323	40.0	1564.0
16883	242043	NaN	3066	67.0	35154.0

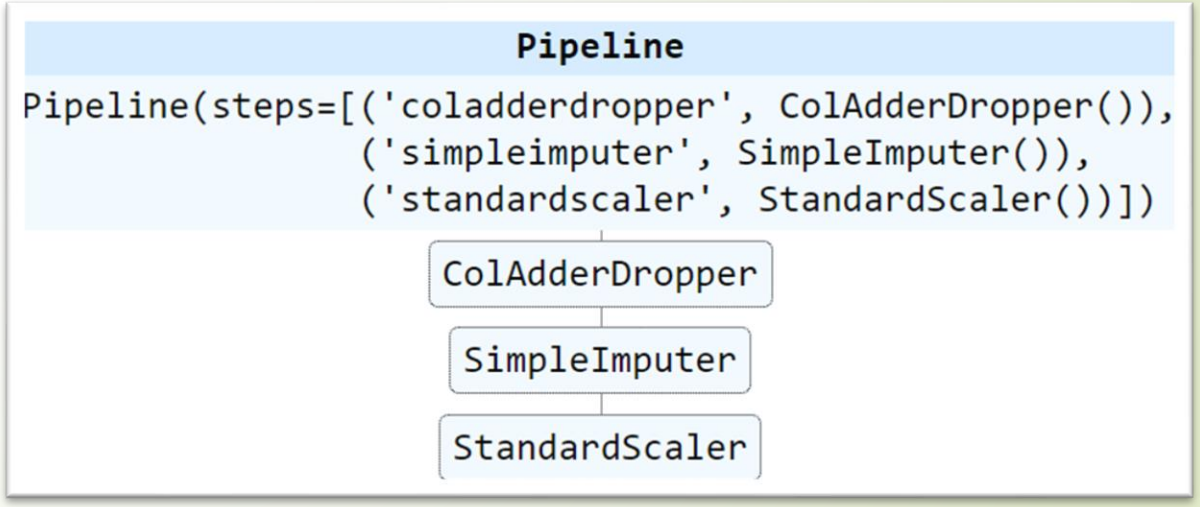
Bộ dữ liệu mới sau khi transform

Tiền xử lý + Mô hình hóa

Preprocess Pipeline

Xây dựng full preprocess pipeline:

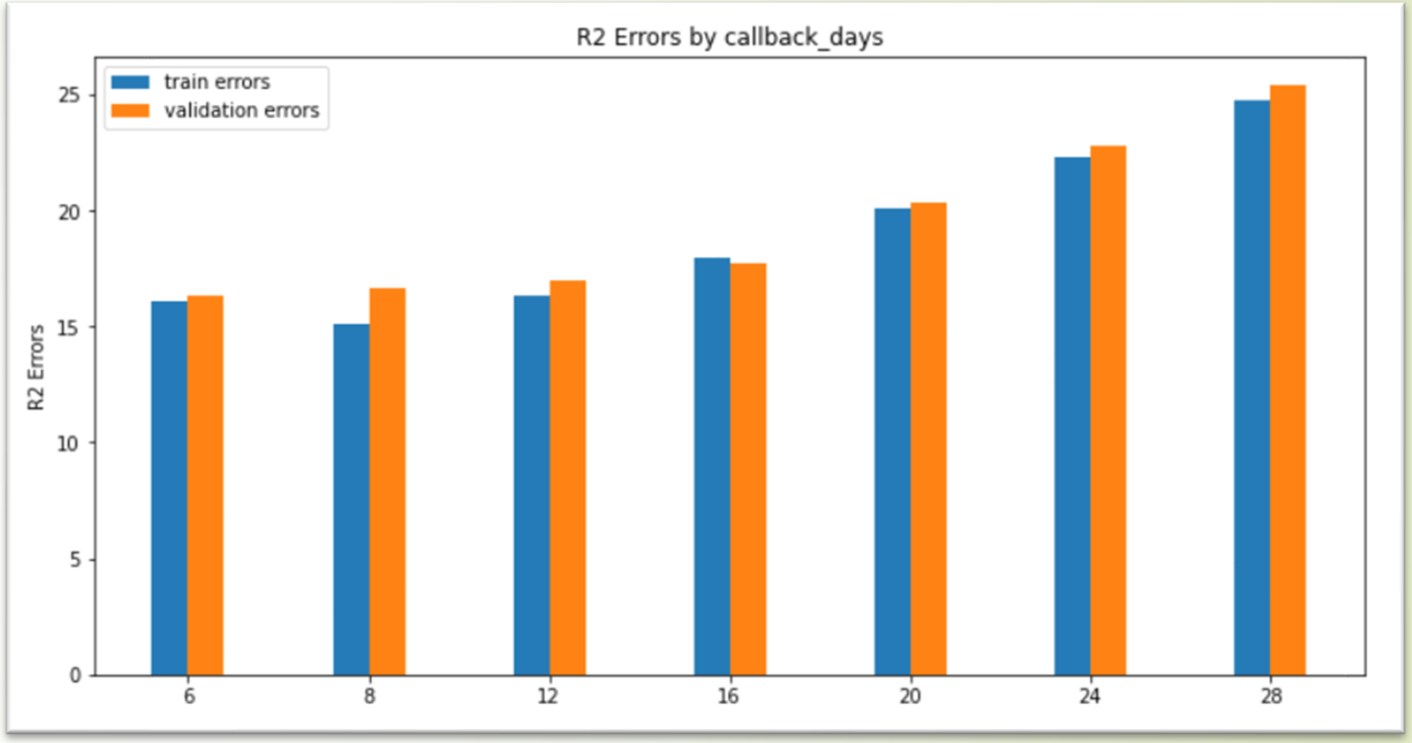
- ColAdderDropper
- SimpleImputer bằng mean
- StandardScaler



Tiền xử lý + Mô hình hóa

Linear Regression

Mô hình hóa sử dụng LinearRegression



Sự thay đổi độ lỗi trên 2 tập train và validation khi thay đổi siêu tham số

Tiền xử lý + Mô hình hóa

Multilayer Perceptron

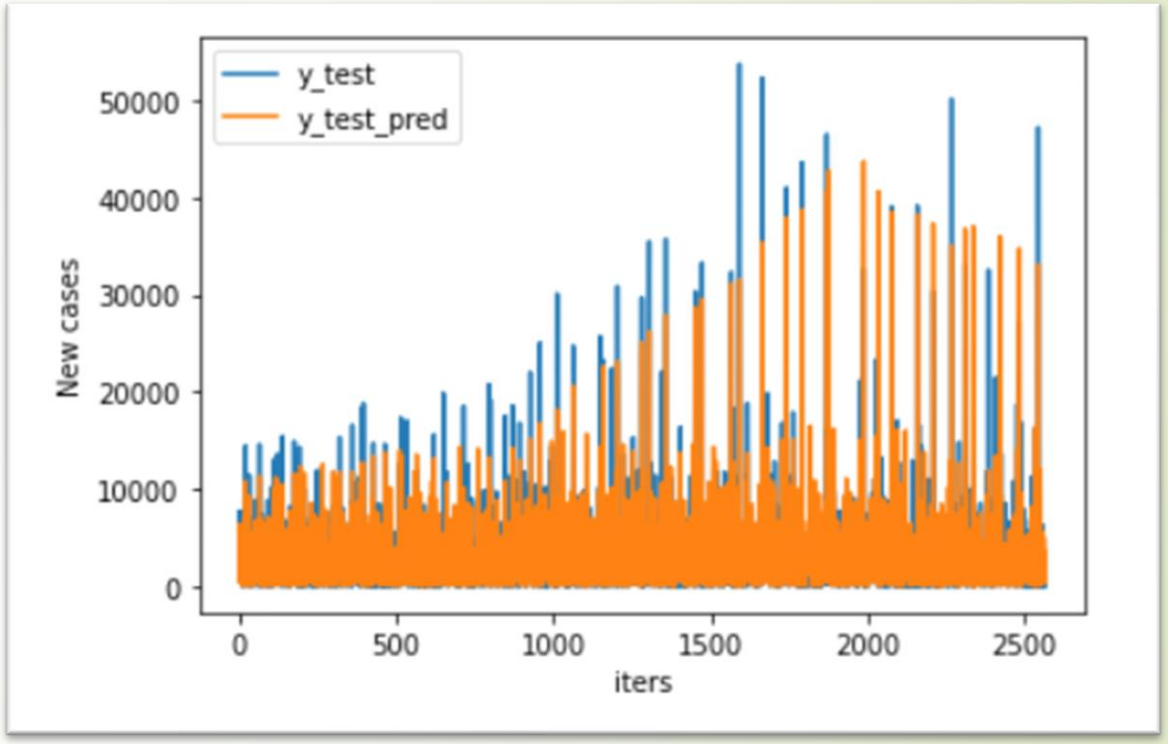
Mô hình hóa sử dụng MLPRegressor

- ❑ Hidden layer sizes: 15
- ❑ Hàm cực tiểu: adam
- ❑ Hàm kích hoạt: relu

Sự thay đổi độ lỗi trên 2 tập train và validation khi thay đổi siêu tham số

Trực quan hóa dữ liệu

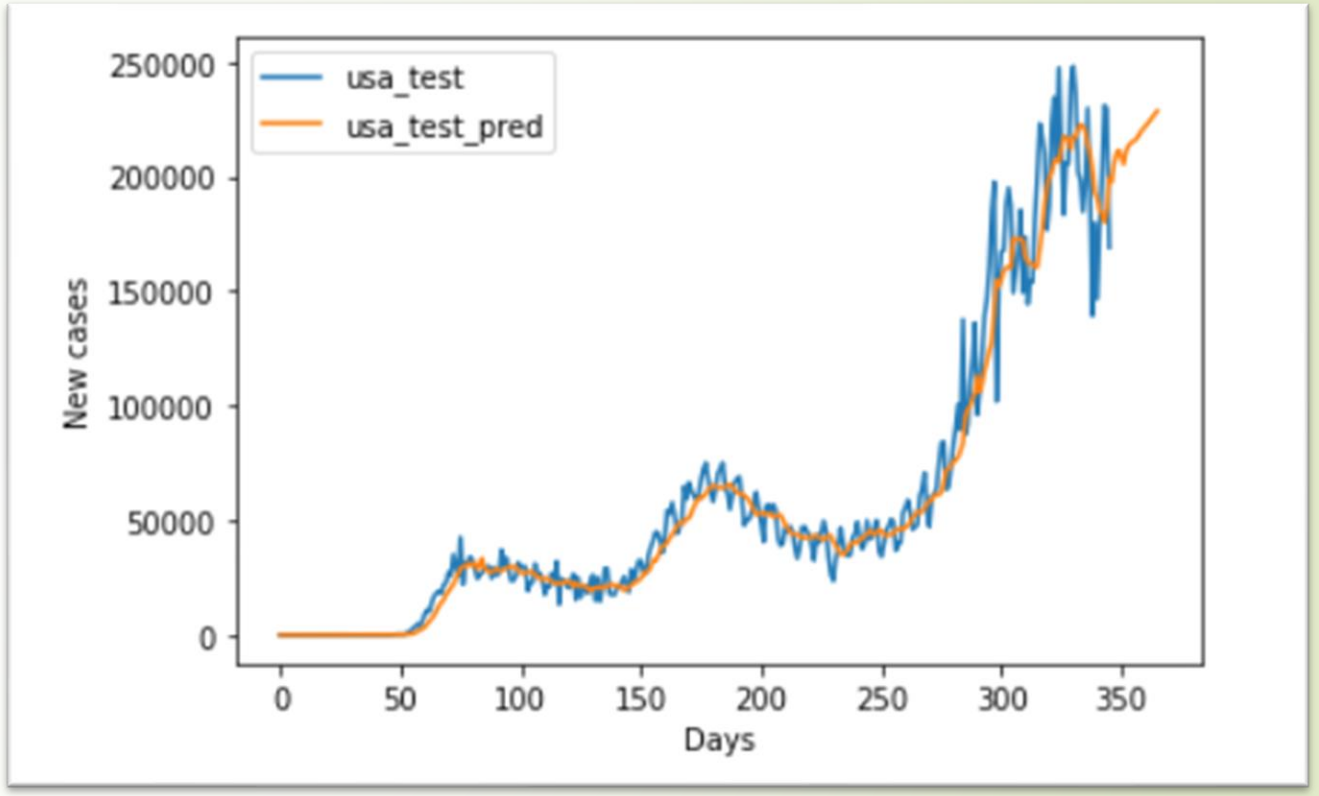
Phân bố kết quả



Trực quan kết quả dự đoán và kết quả thực tế đối với các dòng dữ liệu trong bộ test

Trực quan hóa dữ liệu

Dự đoán cả nước Mỹ



Dự đoán số ca mắc mới mỗi ngày của cả nước Mỹ (kèm theo là dự đoán số ca trong tương lai)

Nhìn lại quá trình làm đồ án

Khó khăn

- Nguồn và độ chính xác của dữ liệu thô.
- Giới hạn thời gian (mô hình hóa mất nhiều thời gian).
- Khó lựa chọn mô hình phù hợp.
- Khó làm việc với Jupyter file trên Github.

Nhìn lại quá trình làm đồ án

Những thứ học được

- Các bước làm hoàn chỉnh trong việc ứng dụng quy trình Khoa học dữ liệu vào bài toán thực tế.
- Các kiến thức, kinh nhiệm về việc lựa chọn siêu tham số và mô hình phù hợp.
- Cách nhận xét và trình bày mô hình tìm được.
- Làm việc nhóm.
- Làm việc với các tính năng nâng cao trên Github.

Cải thiện trong tương lai

Nếu có thêm thêm thời gian:

- Tìm hiểu thêm các mô hình khác chất lượng hơn
- Phân tích và nhận xét kĩ hơn về bài làm
- Tìm thêm các nguồn dữ liệu mới để mô hình hoạt động tốt trên nhiều khu vực hơn

Tài liệu tham khảo

Nguồn dữ liệu và mô tả – Trung tâm Kiểm soát Bệnh tật CDC tại Mỹ

Quy trình Khoa học dữ liệu – Bài tập 3 – Thầy Trần Trung Kiên

Tìm hiểu mô hình MLPRegressor - Sklearn

Ý tưởng hệ số tương quan – Nhóm 14

The End.