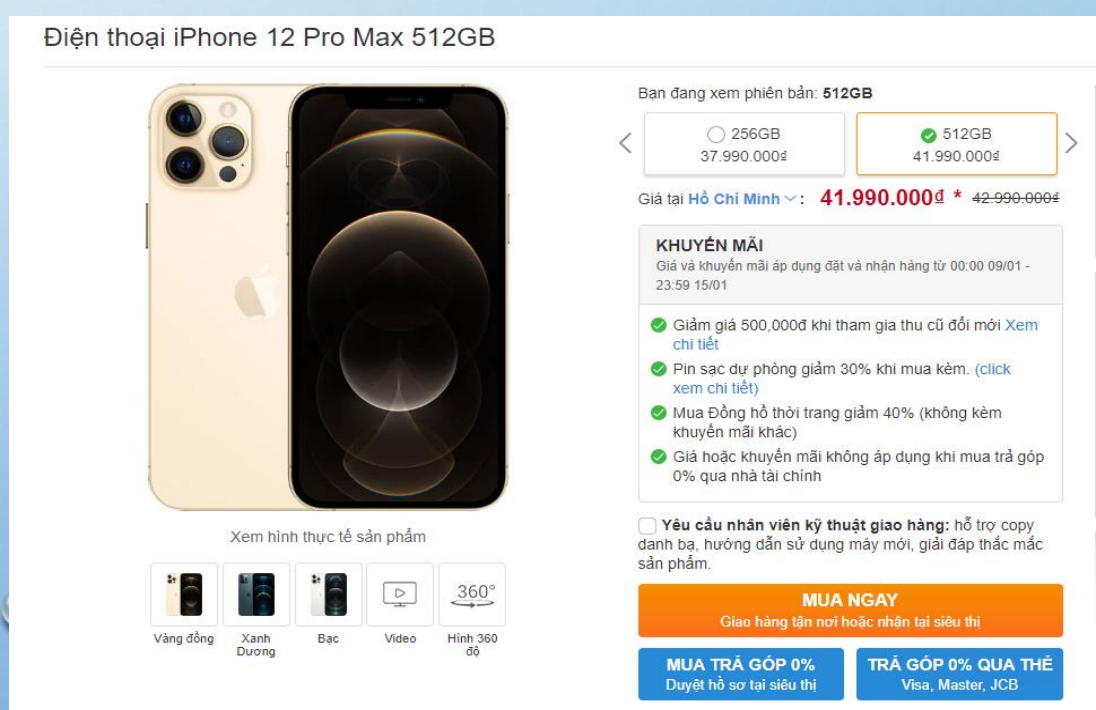
ĐỒ ÁN CUỐI KÌ
DỰ ĐOÁN MỨC GIÁ ĐIỆN THOẠI TỪ TẬP DỮ LIỆU GIÁ ĐIỆN THOẠI Ở VIỆT NAM.

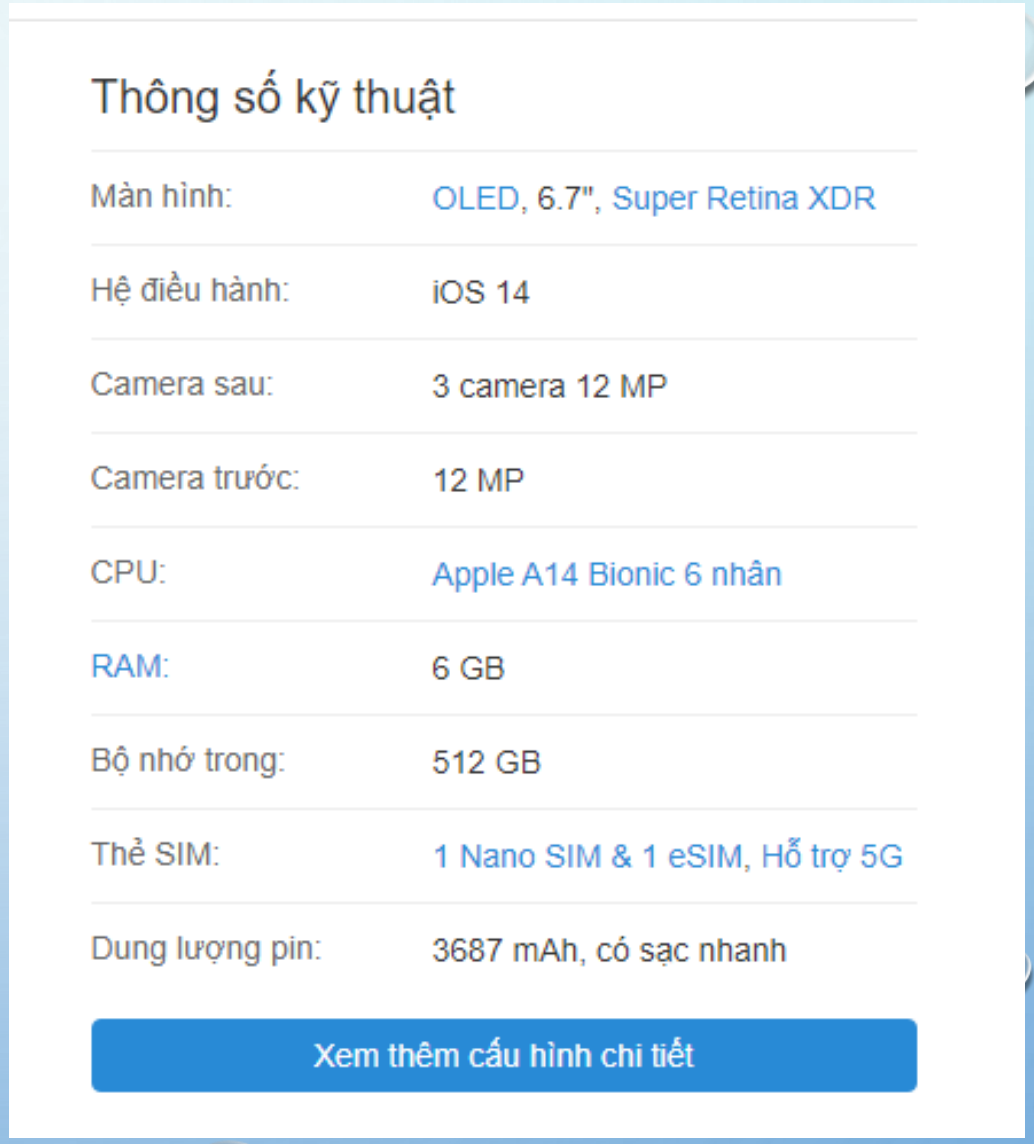
- GIÁO VIÊN: TRẦN TRUNG KIÊN

- SINH VIÊN THỰC HIỆN:
- 18120626 – ĐẶNG QUANG TRƯỜNG
- 18120507 – TRƯƠNG CÔNG PHU



❖NỘI DUNG:

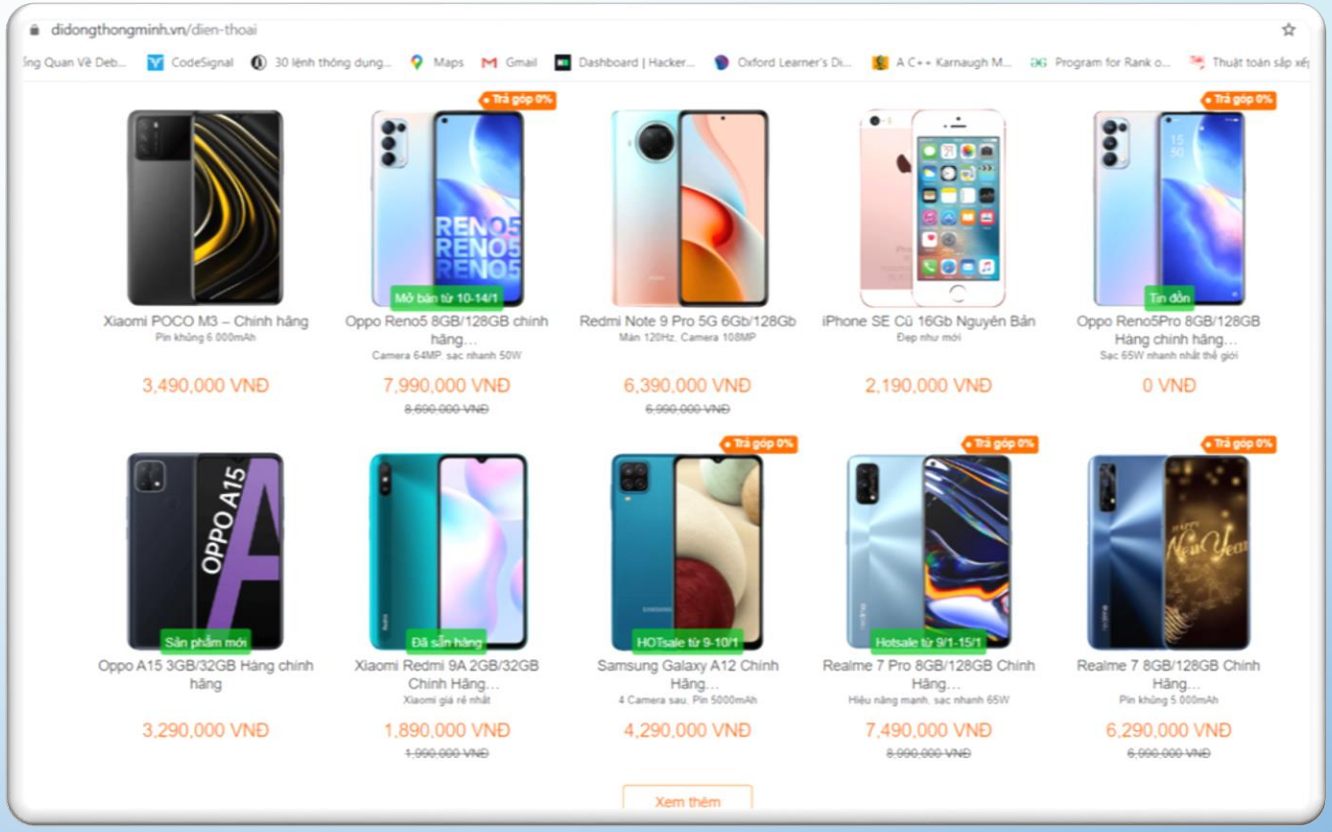
- ✓ Giới thiệu
- ✓ Thu thập dữ liệu
- ✓ Phân tích dữ liệu đưa ra câu hỏi.
- ✓ Tiền xử lý dữ liệu
- ✓ Xây dựng mô hình
- ✓ Đánh giá
- ✓ Tham khảo



GIỚI THIỆU ĐỒ ÁN

- Câu hỏi: Dự đoán tầm giá của điện thoại (giá rẻ, tầm trung, cao cấp, cao cấp(flagship)) dựa vào cấu hình của điện thoại.
- Input: Cấu hình cơ bản của một chiếc điện thoại.
- Output: Cho ra kết quả là điện thoại đó thuộc tầm giá nào. Với các tầm giá sau:
  1. Phân khúc điện thoại giá rẻ: Dưới 6 triệu vnd.
  2. Phân khúc điện thoại tầm trung: Từ 6 triệu vnd đến 10 triệu vnd.
  3. Phân khúc điện thoại cao cấp: Từ 10 triệu vnd đến 18 triệu vnd.
  4. Phân khúc điện thoại cao cấp(flagship): Trên 18 triệu vnd.
- Lợi ích: Phục vụ nhu cầu mua điện thoại của người dùng. Với cấu hình mà người dùng mong muốn có thể đưa ra quyết định giá điện thoại nằm ở khoảng nào.
- Nguồn câu hỏi: Nhóm tự nghĩ.

THU THẬP DỮ LIỆU



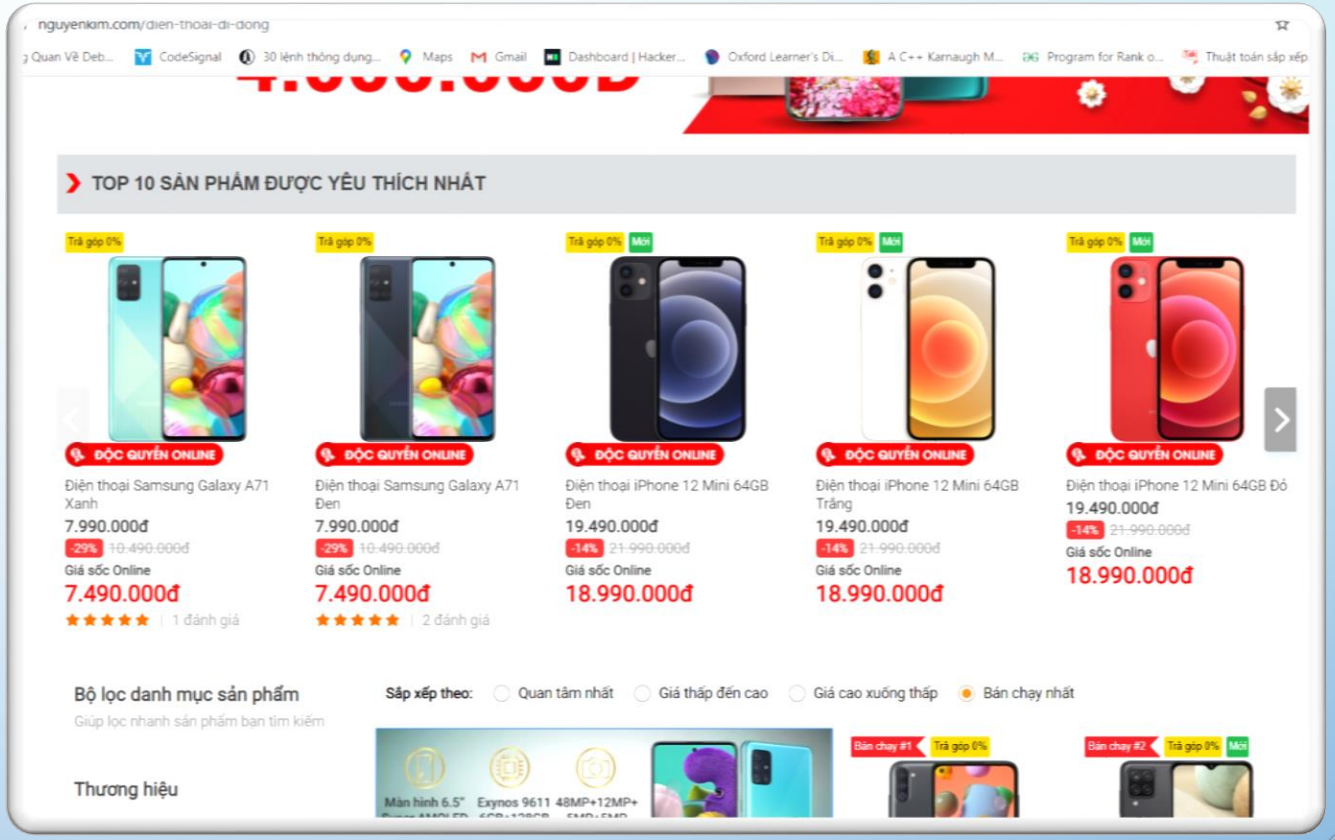
✓ Dữ liệu được thu thập trên 3 trang web bán điện thoại lớn ở Việt Nam:

1. https://didongthongminh.vn/dien-thoai
2. https://www.nguyenkim.com/dien-thoai-di-dong
3. https://www.thegioididong.com/dtdd

THU THẬP DỮ LIỆU

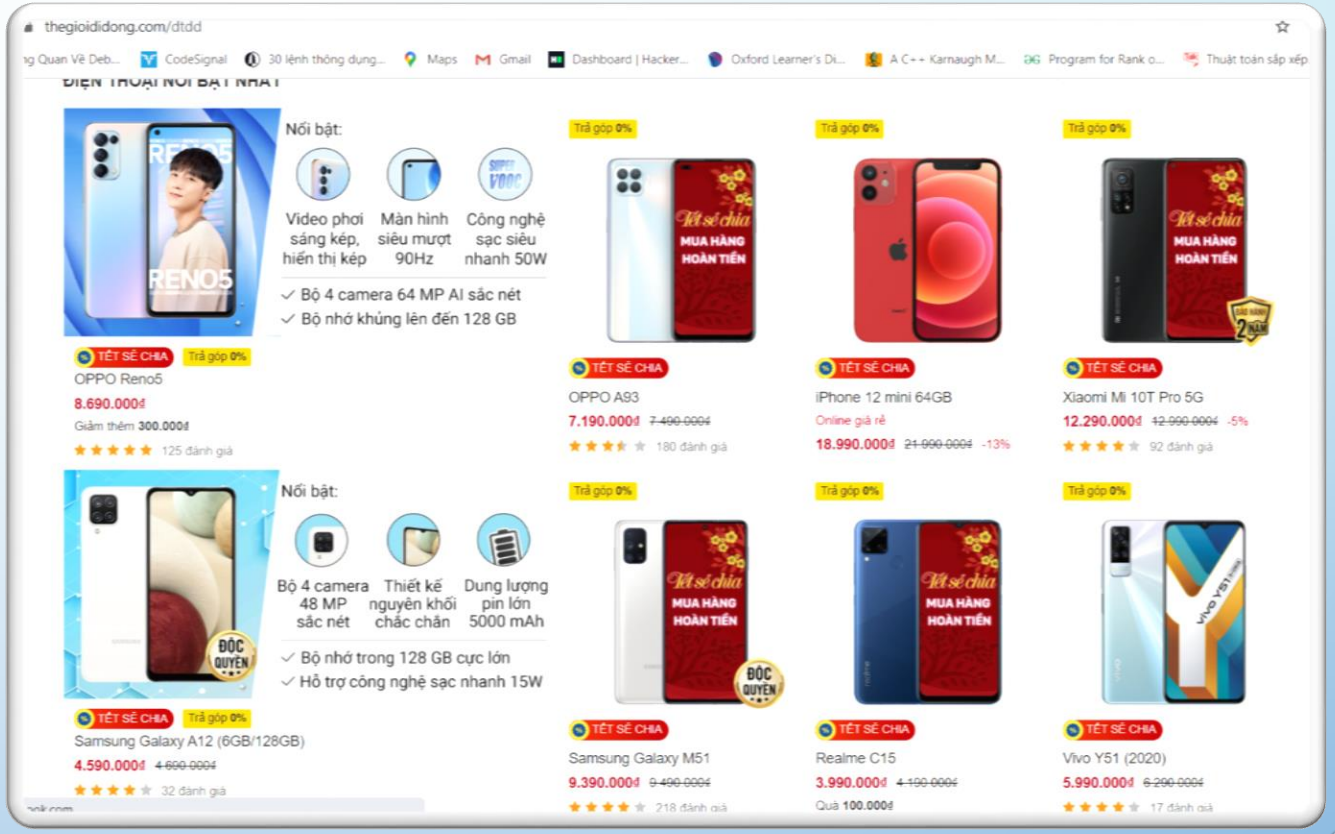
✓ Dữ liệu được thu thập trên 3 trang web bán điện thoại lớn ở Việt Nam:

1. https://didongthongminh.vn/dien-thoai
2. https://www.nguyenkim.com/dien-thoai-di-dong
3. https://www.thegioididong.com/dtdd



THU THẬP DỮ LIỆU

- ✓ Dữ liệu được thu thập trên 3 trang web bán điện thoại lớn ở Việt Nam:
  1. https://didongthongminh.vn/dien-thoai
  2. https://www.nguyenkim.com/dien-thoai-di-dong
  3. https://www.thegioididong.com/dtdd



THU THẬP DỮ LIỆU

- ✓ Phương pháp thu thập: hai phương pháp đã học là parser html và webdriver.
- ✓ Nội dung thu thập: Các giá trị đặc trưng của một chiếc điện thoại.
- ✓ Hạn chế: Vì thu thập trên nhiều trang web khác nhau nên sẽ có dữ liệu lặp, định dạng dữ liệu mỗi trang khác nhau.

THU THẬP DỮ LIỆU

- ✓ Việc thu thập dữ liệu trên 3 trang web này là hợp pháp: đã check robotparser.

Check robotparser

```
]:
1 urls = ['https://www.nguyenkim.com/dien-thoai-di-dong',
2         'https://www.nguyenkim.com/dien-thoai-di-dong',
3         'https://www.thegioididong.com/dtdd']
4 for url in urls:
5     URL_BASE = url
6     parser = robotparser.RobotFileParser()
7     parser.set_url(parse.urljoin(URL_BASE, 'robots.txt'))
8     parser.read()
9     print(url,'check = ', parser.can_fetch('*', url))
```

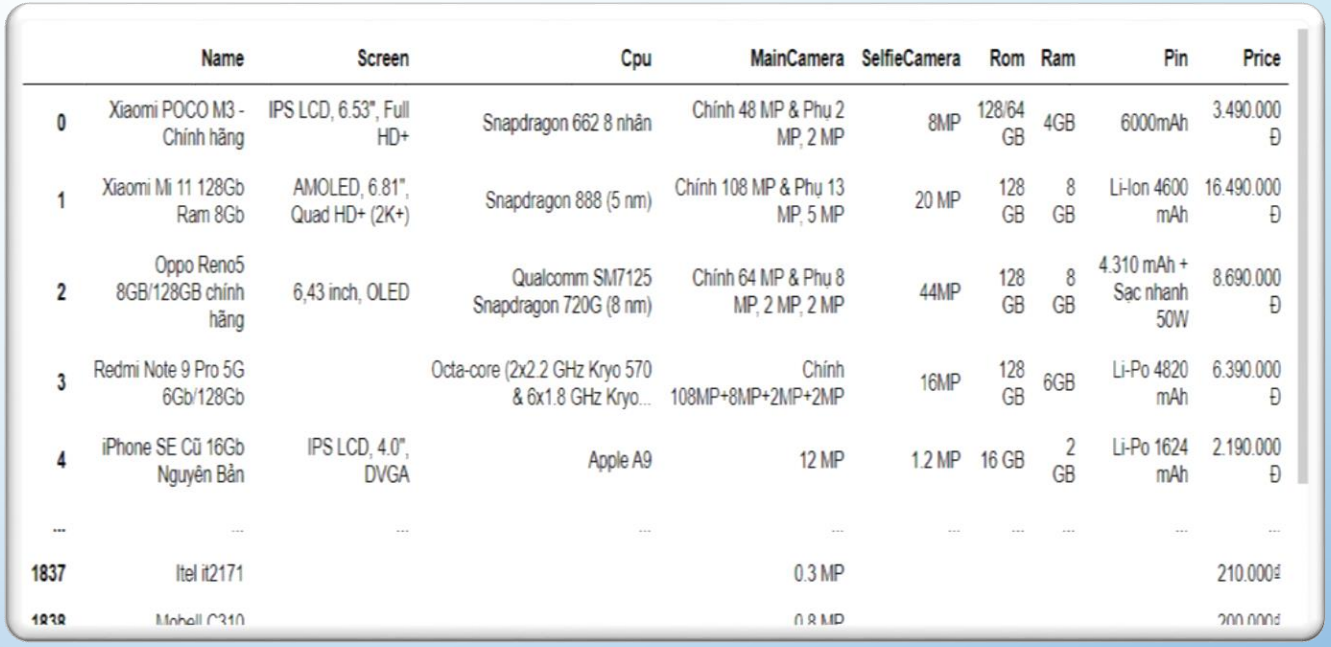
```
https://www.nguyenkim.com/dien-thoai-di-dong check =  True
https://www.nguyenkim.com/dien-thoai-di-dong check =  True
https://www.thegioididong.com/dtdd check =  True
```

Kết quả check ở 3 trang web là True nên ta có thể thu thập dữ liệu hợp pháp từ 3 trang này.

THU THẬP DỮ LIỆU

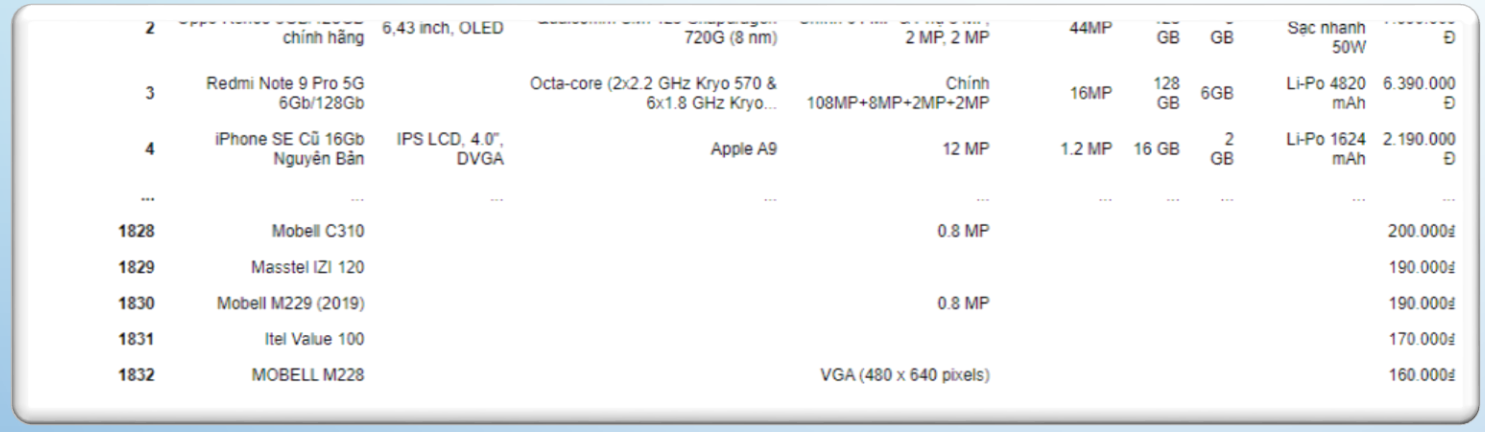
✓ Nội dung thu thập: 1700 dòng dữ liệu đã được loại bỏ trùng lắp.

1. **Name**: Tên điện thoại (có kèm theo hãng sản xuất).
2. **Screen**: Thông tin màn hình.
3. **Cpu**: Cấu hình cpu.
4. **Camera**: Thông tin camera(bao gồm trước và sau).
5. **Ram, Rom**: Thông tin bộ nhớ.
6. **Pin**: Dung lượng pin.
7. **Prince**: Giá bán.



	Name	Screen	Cpu	MainCamera	SelfieCamera	Rom	Ram	Pin	Price
0	Xiaomi POCO M3 - Chính hãng	IPS LCD, 6.53", Full HD+	Snapdragon 662 8 nhân	Chính 48 MP & Phụ 2 MP, 2 MP	8MP	128/64 GB	4GB	6000mAh	3.490.000 Đ
1	Xiaomi Mi 11 128Gb Ram 8Gb	AMOLED, 6.81", Quad HD+ (2K+)	Snapdragon 888 (5 nm)	Chính 108 MP & Phụ 13 MP, 5 MP	20 MP	128 GB	8 GB	Li-Ion 4600 mAh	16.490.000 Đ
2	Oppo Reno5 8GB/128GB chính hãng	6,43 inch, OLED	Qualcomm SM7125 Snapdragon 720G (8 nm)	Chính 64 MP & Phụ 8 MP, 2 MP, 2 MP	44MP	128 GB	8 GB	4.310 mAh + Sạc nhanh 50W	8.690.000 Đ
3	Redmi Note 9 Pro 5G 6Gb/128Gb		Octa-core (2x2.2 GHz Kryo 570 & 6x1.8 GHz Kryo...	Chính 108MP+8MP+2MP+2MP	16MP	128 GB	6GB	Li-Po 4820 mAh	6.390.000 Đ
4	iPhone SE Cũ 16Gb Nguyên Bản	IPS LCD, 4.0", DVGA	Apple A9	12 MP	1.2 MP	16 GB	2 GB	Li-Po 1624 mAh	2.190.000 Đ
...	...	...	...	...	...	...	...	...	...
1837	Itel it2171			0.3 MP					210.000đ
1838	Mobell C310			0.8 MP					200.000đ

TIỀN XỬ LÝ DỮ LIỆU



2	chính hãng	6,43 inch, OLED	720G (8 nm)	2 MP, 2 MP	44MP	GB	GB	Sạc nhanh 50W	Đ
3	Redmi Note 9 Pro 5G 6Gb/128Gb		Octa-core (2x2.2 GHz Kryo 570 & 6x1.8 GHz Kryo...	Chính 108MP+8MP+2MP+2MP	16MP	128 GB	6GB	Li-Po 4820 mAh	6.390.000 Đ
4	iPhone SE Cũ 16Gb Nguyên Bản	IPS LCD, 4.0", DVGA	Apple A9	12 MP	1.2 MP	16 GB	2 GB	Li-Po 1624 mAh	2.190.000 Đ
...	...	...	...	...	...	...	...	...	...
1828	Mobell C310			0.8 MP					200.000₫
1829	Masstel IZI 120								190.000₫
1830	Mobell M229 (2019)			0.8 MP					190.000₫
1831	Itel Value 100								170.000₫
1832	MOBELL M228			VGA (480 x 640 pixels)					160.000₫

- Xóa những dòng bị thiếu hơn 50% thuộc tính (5 thuộc tính)

TIỀN XỬ LÝ DỮ LIỆU

```
In [9]: for idx in df.Price.index:
            price = int(df.Price[idx].replace('.','').replace('Đ','').replace('đ','').replace('₫','').replace(' ',''))
            if price < 6000000:
                df['Price'][idx] = 1#giá rẻ
            elif price <10000000:
                df['Price'][idx] = 2#tầm trung
            elif price < 18000000:
                df['Price'][idx] = 3#cao cấp
            else:
                df['Price'][idx] = 4#cao cấp (flagship)
```

- Phân cấp thuộc tính lớp thành 4 cấp (1, 2, 3, 4)

TIỀN XỬ LÝ DỮ LIỆU

Out[17]:

	Screen	MainCamera	SelfieCamera	Rom	Ram	Pin	PhoneMaker
1414	IPS LCD, 5.7", HD+	5 MP	5 MP	16 GB	1 GB	2800 mAh	OTHERS
1379	IPS LCD, 6.5", HD+	Chính 12 MP & Phụ 2 MP, 2 MP	8 MP	128 GB	4 GB	4230 mAh	OPPO
112	6.4 inches	Triple: 12 MP, f/1.5-2.4, 26mm (wide), 1/2.55"...	8.0 MP	128 GB	8 GB	Li-Ion 4100 mAh	SAMSUNG
636		64MP f1.8 (main) 13MP f2.4(wide) 8MP f/2.4 (te...	Dual 20 MP + 2 MP	64 GB	6 GB	4500mAh	XIAOMI
1312	OLED, 6.1", Super Retina XDR	3 camera 12 MP	12 MP	256 GB	6 GB	2815 mAh, có sạc nhanh	IPHONE
...	...	...	...	...	...	...	...
201	TFT LCD, 6.5", HD+	Chính 13 MP & Phụ 2 MP, 2 MP	5MP	32 GB	3GB	Li-Po 5000 mAh	SAMSUNG
990	Super LCD 3, 5.2", QuadHD (2K)	20 MP	4.0 MP	32 GB	3 GB	Li-Po 2840 mAh	OTHERS
595	5.5 inches, S-CG Silicon LCDFull HD 1080	13.1 MP	1.2 CMOS	16 GB	2 GB	Li-Ion 2610 mAh	OTHERS

- Thay thế thuộc tính Name bằng thuộc tính PhoneMaker (hãng điện thoại) bằng cách chỉ lấy num_top_pmakers giá trị xuất hiện nhiều nhất

TIỀN XỬ LÝ DỮ LIỆU

- Xóa những dòng có thuộc tính lớp bị thiếu (thuộc tính Price)
- Bỏ cột CPU vì cột này có rất nhiều giá trị khác nhau, nếu chuyển sang dạng số bằng phương pháp one_hot sẽ làm tăng số lượng cột rất nhiều

TIỀN XỬ LÝ DỮ LIỆU

Vì dữ liệu thu thập ở một số cột còn chứa chuỗi chữ nên cần tách lấy số và chuyển cột về dạng số (Screen, MainCamera, SelfieCamera, Rom, Ram, Pin)

Điền các giá trị thiếu của cột PhoneMaker bằng giá trị mode

Điền các giá trị thiếu của các cột số bằng giá trị mean

Chuyển cột PhoneMaker sang dạng số bằng phương pháp mã hóa one-hot

TIỀN XỬ LÝ DỮ LIỆ

Dữ liệu sau khi tiền xử lý

	PhoneMaker	Screen	MainCamera	SelfieCamera	Rom	Ram	
1414	OTHERS	5.7	5.0	5.0	16.0	1.0	2.800
1379	OPPO	6.5	12.0	8.0	128.0	4.0	4.230
112	SAMSUNG	6.4	12.0	8.0	128.0	8.0	4.100
636	XIAOMI	NaN	64.0	20.0	64.0	6.0	4.500
1312	IPHONE	6.1	3.0	12.0	256.0	6.0	2.815
...	...	...	...	...	...	...	...
201	SAMSUNG	6.5	13.0	5.0	32.0	3.0	5.000
990	OTHERS	3.0	20.0	4.0	32.0	3.0	2.840
595	OTHERS	5.5	13.0	1.2	16.0	2.0	2.610
229	IPHONE	4.7	8.0	1.2	16.0	1.0	1.810
745	OTHERS	4.8	5.0	NaN	8.0	768.0	1.230

MÔ HÌNH HÓA DỮ LIỆU

- Tách các tập X, y
- Tách các tập train, validation, test (60%, 20%, 20%)

(Việc tiền xử lý dữ liệu sẽ được thực hiện sau khi tách các tập)

- Tạo full_pipeline gồm các bước: xử lý dữ liệu, mô hình hóa dữ liệu
- Test các mô hình mô hình hóa dữ liệu với các tham số của mô hình để tìm ra mô hình tối ưu nhất
- Áp dụng mô hình tối ưu vào full_pipeline với tập dữ liệu train_X, train_y

MÔ HÌNH HÓA DỮ LIỆU – MÔ HÌNH MLPCLASSIFIER

❖Huấn luyện mô hình

➢Lựa chọn mô hình: Mục tiêu là phân nhóm dữ liệu nên nhóm sử dụng mô hình MLPClassifier do scikit-learn hổ trợ.

- ➢Sử dụng nhiều giá trị num_top_pmakers của thuộc tính Phone_makers
- ➢Thay đổi các giá trị alpha
- ➢Sử dụng nhiều hàm kích hoạt khác nhau.

Tìm ra các giá trị tốt nhất cho mô hình (độ lỗi trên tập validation là nhỏ nhất).

MÔ HÌNH HÓA DỮ LIỆU – MÔ HÌNH MLPCLASSIFIER

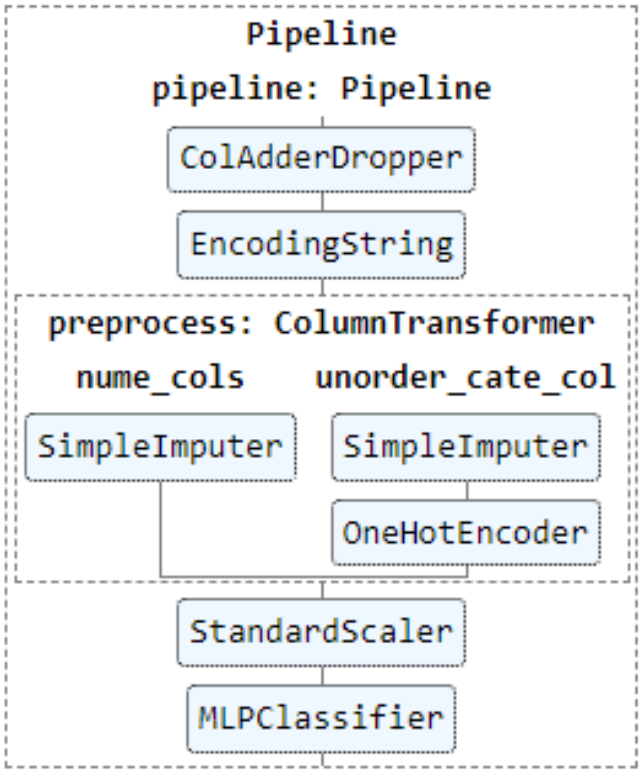
❖Huấn luyện mô hình

➢Quá trình huấn luyện:

- ➢Tạo pipeline để tạo ra một đường ống liên tục trong quá trình xử lý dữ liệu và mô hình
- ➢Huấn luyện mô hình với tập dữ liệu train_x, train_y
- ➢Đánh giá độ chính xác của mô hình trên tập dữ liệu test

MÔ HÌNH HÓA DỮ LIỆU – MÔ HÌNH MLPCLASSIFIER

Out[33]:



MÔ HÌNH HÓA DỮ LIỆU – MÔ HÌNH MLPCLASSIFIER

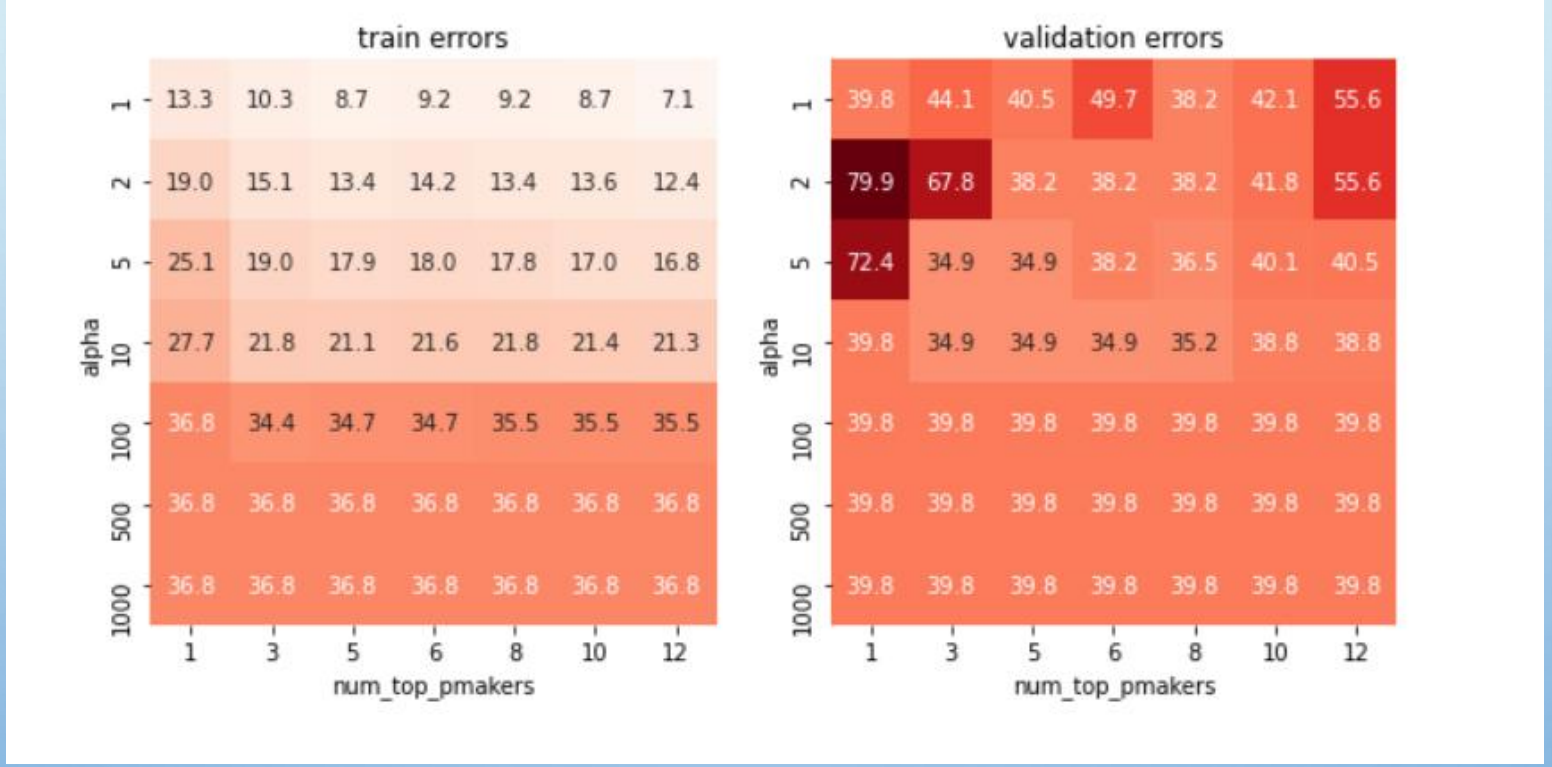
- Thử nghiệm mô hình với MLPClassifier: Tạo một full_pipeline_mlpclassifier với các tham số:
  - ✓ Hàm kích hoạt: activation = 'tanh'
  - ✓ Các lớp ẩn: Sử dụng 1 lớp ẩn hidden_layer_sizes=(20)
  - ✓ Max_iter = 15000

```
full_pipeline_mlpclassifier = make_pipeline(
    (preprocess_pipeline),
    (MLPClassifier(hidden_layer_sizes=(20),activation= 'tanh', solver='lbfgs', random_state=0, max_iter=15000))

)
```

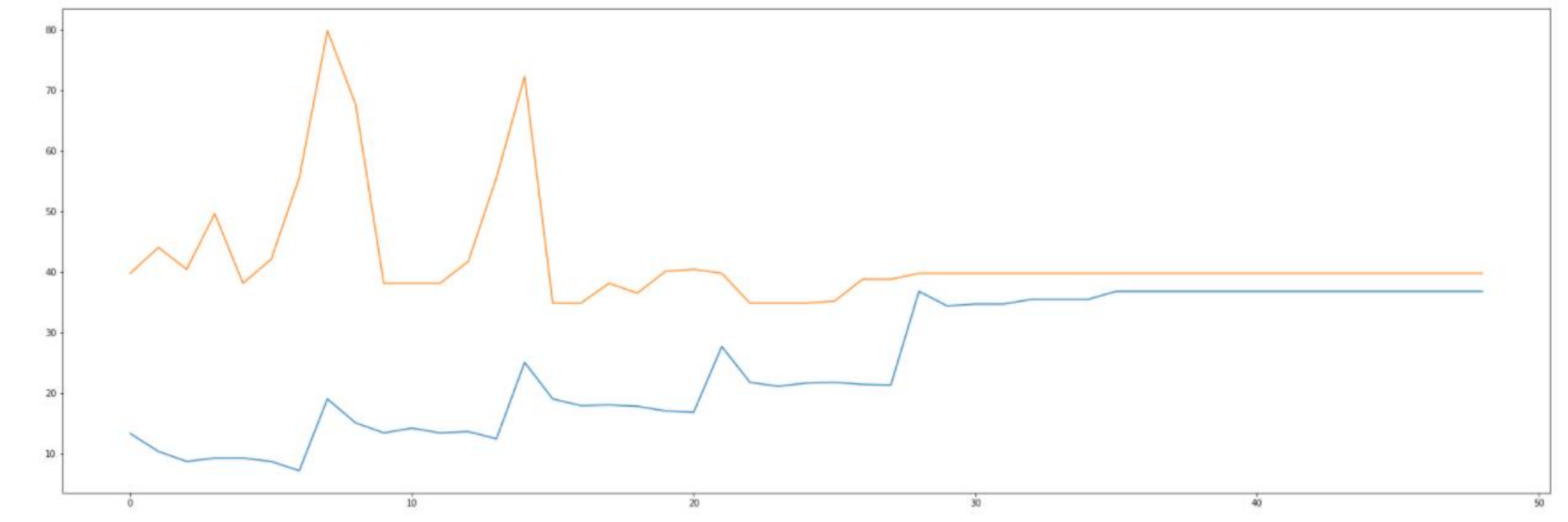
- Sử dụng nhiều giá trị khác nhau của tham số alpha và num_top_pmakers để dự đoán và tìm ra mô hình tốt nhất.

```
alphas = [1, 2, 5, 10, 100, 500, 1000]
num_top_pmakers_s = [1, 3, 5, 6, 8,10,12]

best_val_err = float('inf')
best_alpha = None
best_num_top_pmakers = None
```

TRỰC QUAN KẾT QUẢ MÔ HÌNH MLPCLASSIFIER TRÊN TẬP TRAIN VÀ VALIDATION VỚI CÁC THAM SỐ KHÁC NHAU



TRỰC QUAN KẾT QUẢ MÔ HÌNH MLPCLASSIFIER TRÊN TẬP TRAIN VÀ VALIDATION VỚI BEST_ALPHA VÀ BEST_NUM_TOP_PMAKERS



DỰ ĐOÁN MÔ HÌNH TRÊN TẬP TRAIN VỚI GIÁ TRỊ TỐT NHẤT CỦA ALPHA VÀ NUM_TOP_PMAKERS

```
best_val_err:  34.86842105263158

best_num_top_pmakers:  6

best_alpha:  10
```

- Mô hình hóa tập dữ liệu train với các giá trị best_alpha và best_num_top_pmakers nhận được từ trên:

```
]:  1 full_pipeline_mlpclassifier.set_params(pipeline__coladderdropper__num_top_pmakers=best_num_top_pmakers,mlpclassifier__alpha
    2 m_mlpclassifier=full_pipeline_mlpclassifier.fit(train_X,train_y)
```

DỰ ĐOÁN KẾT QUẢ TẬP TEST

```
[ ]:  1  # Độ lỗi của mô hình dự đoán trên tập test.
      2  print('Test error: ', (1- m_mlpclassifier.score(test_X,test_y))*100)
```

```
Test error:  36.51315789473685
```

MÔ HÌNH HÓA DỮ LIỆU – MÔ HÌNH LOGISTICREGRESSION

❖Huấn luyện mô hình

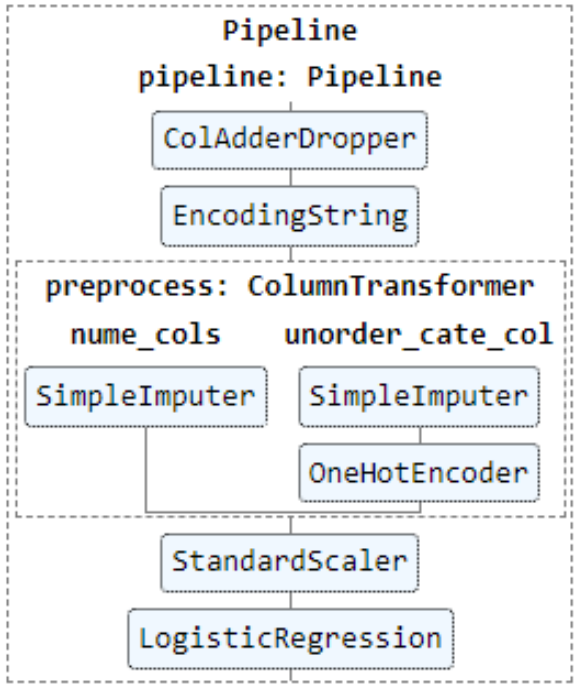
➢Lựa chọn mô hình: Sử dụng mô hình LogisticRegreesion do scikit-learn hỗ trợ.

➢Quá trình huấn luyện:

- ➢Tạo pipeline để tạo ra một đường ống liên tục trong quá trình xử lý dữ liệu và mô hình
- ➢Huấn luyện mô hình với tập dữ liệu train_x, train_y
- ➢Đánh giá độ chính xác của mô hình trên tập dữ liệu test

MÔ HÌNH HÓA DỮ LIỆU

Out[38]:



DỰ ĐOÁN KẾT QUẢ TRÊN TẬP TEST

```
]:  1 # Tính độ lỗi trên tập test
    2 print('Test error: ',(1-m_logisticregression.score(test_X,test_y))*100)
```

```
Test error:  36.51315789473685
```

TỔNG KẾT

❖**Khó khăn:**

+Vì số mẫu dữ liệu từ mỗi trang web nhỏ nên việc thu thập dữ liệu từ nhiều trang web và tiền xử lý mất nhiều thời gian

+Việc đặt câu hỏi trong quá trình làm nhóm có sự thay đổi từ dự đoán giá điện thoại sa ng phân cấp dòng điện thoại

❖**Kiến thức học được trong quá trình làm đồ án:**

+Nắm được nhiều kiến thức hơn về github, jupyter notebook, hiểu hơn các mô hình neural network

TÀI LIỆU THAM KHẢO

- Trực quan hóa dữ liệu và tiền xử lí dữ liệu: File bài tập 3 : Tiền xử lí và mô hình hóa.
- MPLClassifier: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
- LogistcRegression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

CẢM ƠN THẦY
VÀ CÁC BẠN
ĐÃ THEO DÕI !