

Trường Đại học Công nghệ Thông tin
Khoa Khoa học và Kỹ thuật thông tin

DỰ ĐOÁN GIÁ LAPTOP VÀ DESKTOP CŨ TRÊN CHỢ TỐT

Nhóm 6:

10. Trương Thị Thanh Thanh - KHMT
33. Huỳnh Nguyễn Văn Khánh - KHMT
36. Nguyễn Thị Ngọc Nga - KHMT



Nội dung

1 Giới thiệu

3 Phương pháp phân tích

5 Kết luận

2 Mô tả bộ dữ liệu

4 Kết quả phân tích

6 DEMO

01. Giới thiệu

DỰ ĐOÁN
GIÁ MÁY
TÍNH CŨ



Công cụ hỗ trợ

Xây dựng



Lưu trữ



Google Drive

Crawl



BeautifulSoup



02. Mô tả bộ dữ liệu

Bắt đầu

Khảo sát trang web

Xác định vị trí của các thông tin cần crawl (xác định các thẻ chứa thông tin)

Thu thập link của các bài đăng

Truy cập trang danh sách bài đăng

Thu thập link bài đăng

Truy cập trang hiển thị kế tiếp

Thu thập thông tin chi tiết

Truy cập link đã lưu theo danh sách

Thu thập thông tin trên bài đăng

Lưu thông tin

Sau khi thu thập link bài đăng trên 200 trang hiển thị

Lưu trữ dữ liệu thành file .csv

Kết thúc

Dữ liệu thu thập có tổng cộng 8004 dòng và 13 cột, trong đó bao gồm biến mục tiêu (price).

Trong 13 cột này, có 12 cột được lấy trực tiếp từ trang chotot.com, còn một cột (classify - phân loại).

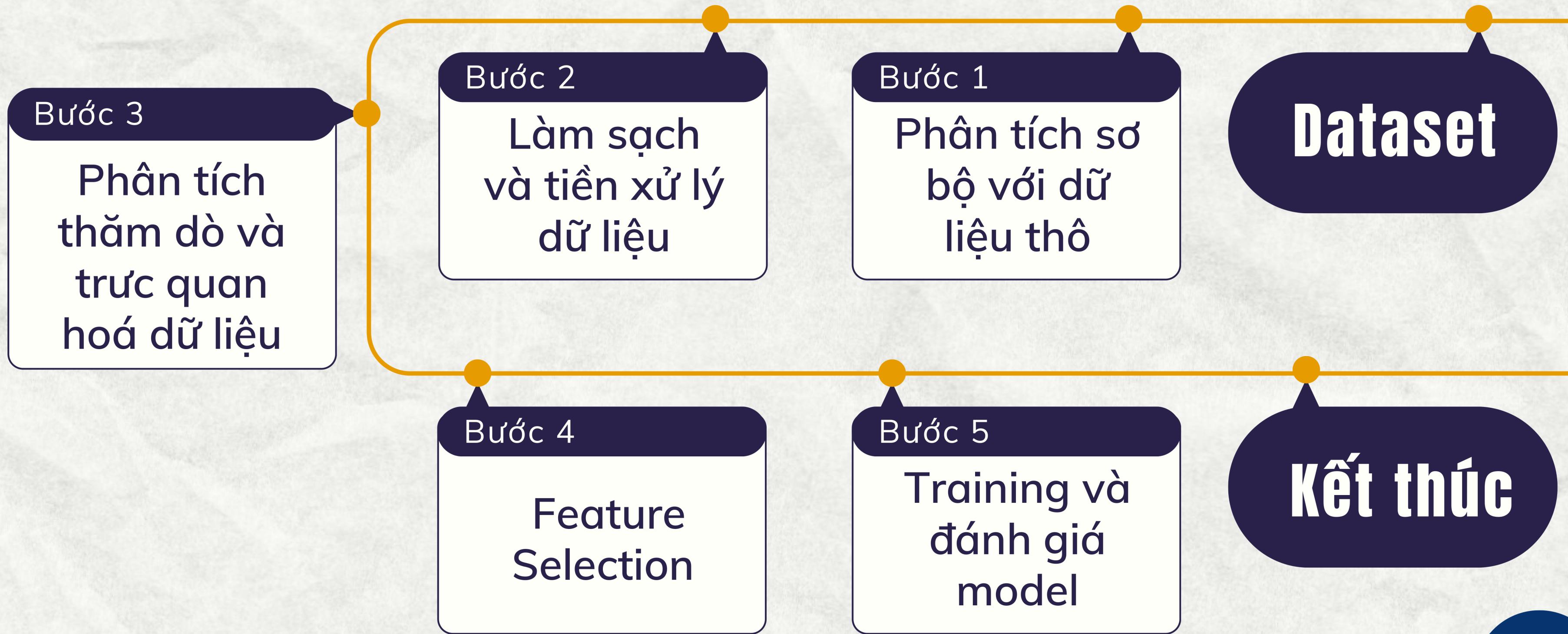


02. Mô tả bộ dữ liệu

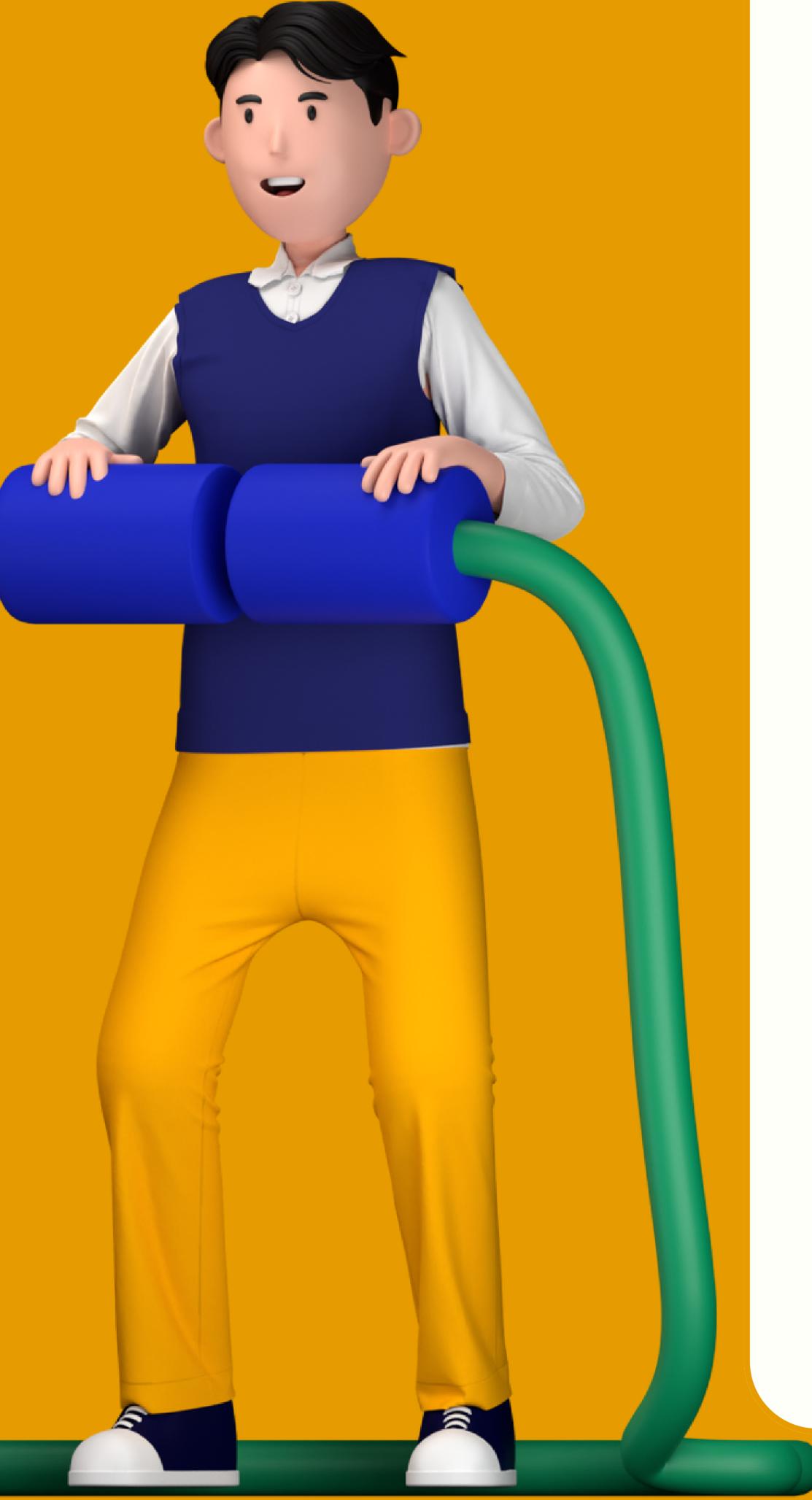
pc_brand	Phân loại	Nhãn hiệu của máy tính. Vd: DELL, Apple, Lenovo, ...
pc_model	Phân loại	Dòng máy. Vd: ThinkPad, ...
elt_condition	Phân loại	Tình trạng sử dụng máy, là máy mới hay đã qua sử dụng, máy đã qua sửa chữa hay chưa.
elt_warranty	Phân loại	Thời gian bảo hành. Vd: 1 tháng, 2 tháng, hết bảo hành, ...
desktop_screen_size	Phân loại	Thông tin về kích cỡ màn hình của laptop dựa trên màn hình tích hợp, trong khi đối với máy tính để bàn, việc có màn hình đi kèm hay không đi kèm phụ thuộc vào từng trường hợp cụ thể.
pc_cpu	Phân loại	Dòng cpu của máy. Vd: Intel Core i5, Intel Core i3, ...
pc_ram	Phân loại	Dung lượng của RAM. Vd: 128 GB, ...
pc_vga	Phân loại	Card đồ họa của máy. Vd: NVIDIA, Onboard, ...
pc_drive_capacity	Phân loại	Dung lượng của ổ cứng.
elt_origin	Phân loại	Xuất xứ của máy. Vd: Việt Nam, ...
usage_information	Phân loại	Thông tin sử dụng (khác với tình trạng sử dụng)
classify	Phân loại	Phân loại là laptop hay máy bàn



03. Phương pháp phân tích



B1. Phân tích sơ bộ



Dữ liệu Thiếu sót nhiều: Một thách thức lớn là sự thiếu sót thông tin trong dữ liệu, đặc biệt là các trường có link không khả dụng. Có trường hợp mà toàn bộ cột đều bị thiếu, điều này đặc biệt đối mặt với những trường hợp mà link đã bị xóa.

Thiếu thông tin quan trọng ở các vị trí ý nghĩa: Phát hiện rằng có một số vị trí quan trọng bị thiếu thông tin, như pc_brand và pc_model, đặc biệt là trong trường hợp của máy tính để bàn ('desktop'). Điều này tạo ra thách thức trong việc xử lý và đánh giá dữ liệu liên quan đến máy tính để bàn.

Dữ liệu Trùng lặp: Bộ dữ liệu chứa nhiều dòng trùng lặp, yêu cầu thực hiện quy trình loại bỏ để đảm bảo tính độc lập và tin cậy của dữ liệu. Cụ thể, có 957 dòng dữ liệu bị trùng lặp.

Giá trị Nhiều trong pc_model và pc_cpu: Giá trị "Dòng khác" hoặc "Dòng Khác" thường xuất hiện trong pc_model, gây nhiễu và khó khăn trong quá trình phân tích. Tương tự, giá trị "Khác" trong pc_cpu ảnh hưởng đáng kể đối với các mô hình đặc biệt với pc_brand là "Apple".

B2. Làm sạch

8004

7356

7355

6636

6636

Số dòng dữ liệu thu thập được

Xóa bỏ các dữ liệu trùng lặp

Loại bỏ những hàng thiếu giá trị "price"

Loại bỏ những hàng thiếu cả 3 giá trị "pc_cpu",
"pc_ram", và "pc_drive_capacity"

Thay thế giá trị nhiễu và khuyết còn lại

B2. Tiền xử lí dữ liệu

#	Column	Non-Null Count	Dtype
0	pc_brand	6636 non-null	object
1	pc_model	6636 non-null	object
2	elt_condition	6636 non-null	object
3	elt_warranty	6636 non-null	object
4	desktop_screen_size	6636 non-null	object
5	pc_cpu	6636 non-null	object
6	pc_ram	6636 non-null	object
7	pc_vga	6636 non-null	object
8	pc_drive_capacity	6636 non-null	object
9	elt_origin	6636 non-null	object
10	classify	6636 non-null	object
11	price	6636 non-null	object
12	pc_cpu_label	6636 non-null	object
dtypes: object(13)			



#	Column	Non-Null Count	Dtype
0	pc_brand	6636 non-null	int64
1	pc_model	6636 non-null	int64
2	elt_condition	6636 non-null	int64
3	elt_warranty	6636 non-null	int64
4	desktop_screen_size	6636 non-null	int64
5	pc_cpu	6636 non-null	int64
6	pc_ram	6636 non-null	int64
7	pc_vga	6636 non-null	int64
8	pc_drive_capacity	6636 non-null	int64
9	elt_origin	6636 non-null	int64
10	classify	6636 non-null	int64
11	price (VND)	6636 non-null	float64
12	pc_cpu_label	6636 non-null	int64
dtypes: float64(1), int64(12)			

Xoá bỏ cột usage_information ra khỏi bộ dữ liệu vì usage_information chỉ chứa duy nhất 1 giá trị.

Thêm cột "pc_cpu_label" để có thể phân tích sự ảnh hưởng của loại cpu đối với "price".

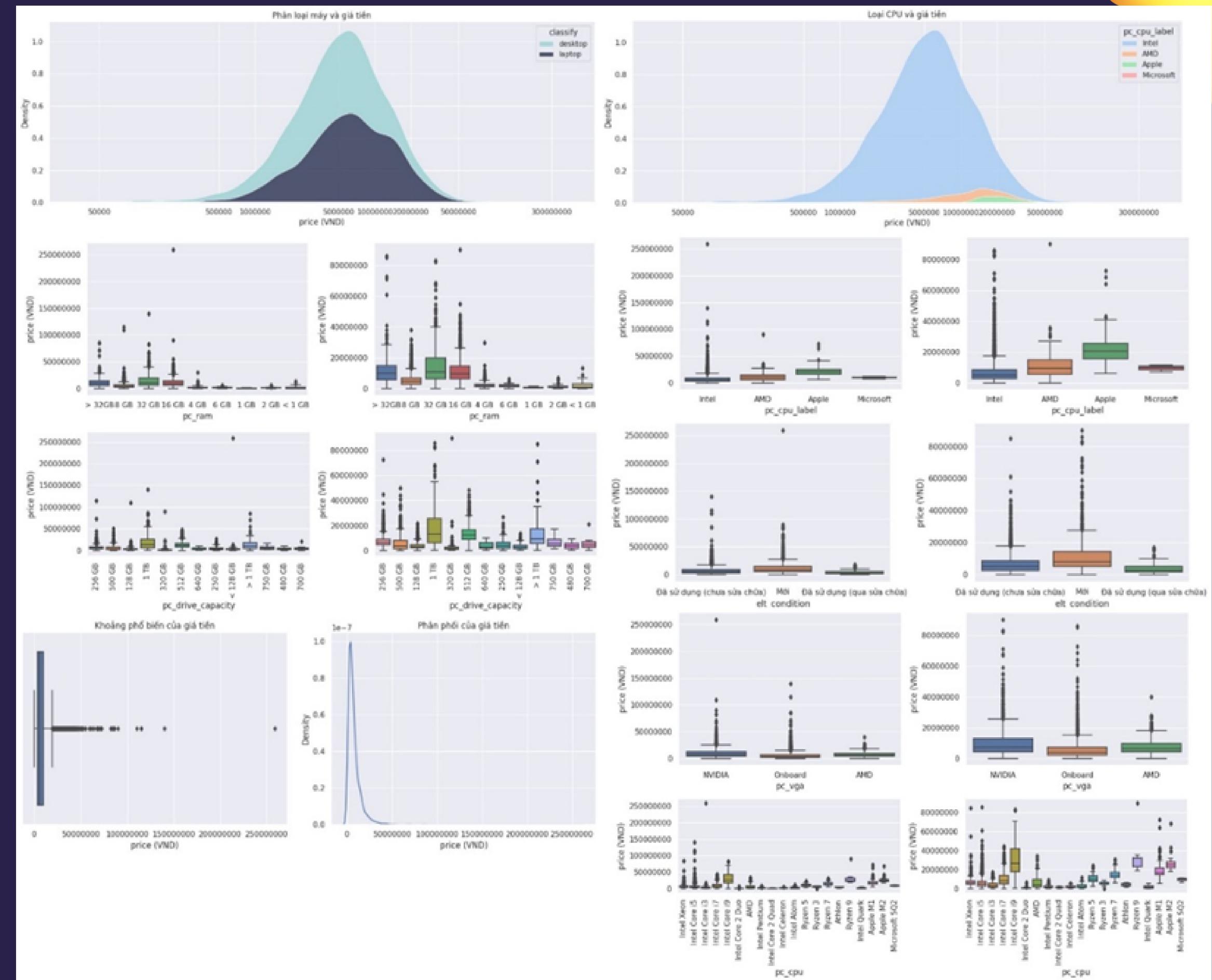
Label encoding cho các cột dữ liệu còn lại, nhằm chuyển đổi từ kiểu dữ liệu "object" sang dạng "int64"

Với tham số price chỉ tiến hành chuyển đổi từ kiểu dữ liệu "object" sang dạng "float64"



B3. Phân tích thăm dò và trực quan hóa dữ liệu

Kết quả thăm dò dữ liệu:



B3. Phân tích thăm dò và trực quan hóa dữ liệu

Bảng độ ảnh hưởng của các biến phân loại (sử dụng ANOVA):

Variable name	F-test	P-value
pc_ram	236.886649	0.000000e+00
elt_condition	189.865192	6.553893e-81
pc_drive_capacity	173.238715	0.000000e+00
pc_cpu_label	171.129723	7.332339e-107
pc_vga	170.881872	4.321303e-73
pc_cpu	120.844553	0.000000e+00
classify	92.396042	9.826792e-22
elt_warranty	82.308319	3.730625e-115
desktop_screen_size	40.420181	1.725495e-63
pc_brand	25.754897	8.159431e-71
elt_origin	23.521912	5.470760e-40
pc_model	16.620947	8.723770e-249

Dựa vào bảng trên, ta thấy giá máy tính có thể bị chi phối mạnh bởi độ lớn của RAM, tình trạng sử dụng, độ lớn bộ nhớ của ổ cứng, loại cpu, dòng cpu và card màn hình

B4. Feature Selection



Loại bỏ một số dòng (outliers) với mức giá lớn hơn 65 triệu VND.

Sử dụng 12 biến để huấn luyện mô hình, đó là: pc_brand (nhãn hiệu), pc_model (dòng máy), elt_condition (tình trạng sử dụng), elt_warranty (thời gian bảo hành), desktop_screen_size (kích thước màn hình), pc_cpu (dòng CPU), pc_ram (RAM), pc_vga (card màn hình), pc_drive_capacity (ổ cứng), elt_origin (xuất xứ), classify (loại máy), pc_cpu_label (loại CPU).

B5. Training và đánh giá model



Trong số bốn biến dạng số, chỉ sử dụng hai biến có sự tương quan với biến mục tiêu (price) cho bước huấn luyện mô hình và loại bỏ một số dòng (outliers) với mức giá lớn hơn 65 triệu VND

Thử nghiệm trên 8 thuật toán máy học khác nhau, bao gồm Histogram Gradient Boosting, XGBoost, Random Forest, KNN, Decision Tree, Linear Regression, SVM và Neural Networks.

Các thuật toán và mô hình sẽ được đánh giá hiệu suất dựa trên ba chỉ số quan trọng sau đây: R2 score, Mean Absolute Error và Mean Squared Error.

Giữ nguyên cài đặt mặc định được sử dụng để đánh giá hiệu suất ban đầu của các mô hình mà không thay đổi tham số.

Lựa chọn mô hình tốt nhất để tiếp tục để thực hiện quá trình fine-tune trên mô hình đã chọn để tìm ra các tham số phù hợp nhất, nhằm đạt được kết quả tốt nhất có thể.

04. Kết quả phân tích



	model	r2	mse	mae
0	Histogram Gradient Boosting	0.709438	1.214390e+13	2.247683e+06
1	XGBoost	0.704299	1.235868e+13	2.246087e+06
2	Random Forest	0.642799	1.492903e+13	2.440168e+06
3	KNN	0.545041	1.874275e+13	2.653571e+06
4	multi_mmscale	0.261507	3.086497e+13	3.723290e+06
5	multi_stscale	0.261507	3.086497e+13	3.723290e+06
6	multi	0.261507	3.086497e+13	3.723290e+06
7	Decision Tree	0.244301	3.158409e+13	3.116065e+06
8	multi_norm	0.221013	3.255738e+13	3.879834e+06
9	SVR	-0.050502	4.390522e+13	4.175250e+06
10	Neural Network	-1.032693	8.495543e+13	6.558531e+06

05. Kết luận

- Các biến được sử dụng: pc_brand (nhãn hiệu), pc_model (dòng máy), elt_condition (tình trạng sử dụng), elt_warranty (thời gian bảo hành), desktop_screen_size (kích thước màn hình), pc_cpu (dòng CPU), pc_ram (RAM), pc_vga (card màn hình) pc_drive_capacity (ổ cứng), elt_origin (xuất xứ), classify (loại máy), pc_cpu_label (loại CPU).
- Sau khi tiến hành huấn luyện mô hình, thu được kết quả khá khả quan với R2 là 0.71, và với mỗi máy, mô hình dự đoán ra kết quả chênh lệch khoảng 2 triệu 300 VND (MAE là 2301025.59).



DEMO

Q&A

**Cảm ơn các bạn
đã lắng nghe**