Báo cáo phương pháp xử lý dữ liệu

Chú thích:

- Nội dung: Báo cáo trình bày các phương pháp xử lý dữ liệu gồm Data Ingestion, Parsing, Cleaning và Preprocessing, cùng những vấn đề phát sinh và cách khắc phục
- Người thực hiện: Nguyễn Thị Minh Tiến - Đoàn Thị Ngọc Anh

I. Data Parsing

4 phương pháp gồm:

1. **Regex:** được sử dụng cho các cột có cấu trúc sẵn, logic dễ diễn giải bằng Python code

Cột ban đầu	Ví dụ	Các cột sau khi xử lý	Đơn vị
dimensions	Dày: 1,13cm - Chiều dài: 30,41cm Chiều rộng: 21,5cm	laptop_height laptop_depth laptop_width	millimeter(mm)
battery	4 cell - 57 Wh	battery_capacity	watt-hour (wh)
laptop_loai_ram	DDR5 5200MHz	ram_type ram_speed	megahertz (mhz)
laptop_so_khe_ram	2 khe (2 x 8GB, máy nguyên bản 8GB, được tặng 8GB, nâng cấp tối đa 32GB)	ram_slot storage_max_support	gigabyte (gb)
o_cung_laptop	512GB PCIe 4.0 NVMe M.2 SSD (2 khe cắm M.2 hỗ trợ SATA hoặc NVMe, Nâng cấp tối đa 1TB)	laptop_storage storage_max_support	gigabyte (gb)
vga	NVIDIA GeForce RTX 2050 4 GB GDDR6 VRAM	vga_type vga_brand vga_vram	gigabyte (gb)
display_resolution	1920 x 1080 pixels (FullHD)	display_height display_width display_type	(pixels)

2. **LLMs:** được sử dụng cho các cột có câu trúc phúc tạp, không thể sử lý bằng logic Python code

- Mô hình sử dụng: Llama 3.3
- Phương pháp: Few-shot Prompting để trích xuất dữ liệu cần thiết
- Prompt:

prompt = '''

Bạn là một trợ lý chuyên gia về trích xuất thông tin CPU laptop. Nhiệm vụ của bạn là phân tích cú pháp (parse) các chuỗi mô tả CPU thô và trích xuất thông tin cấu trúc chi tiết vào các trường được cung cấp.\n

Với **MỖI** chuỗi mô tả CPU được cung cấp trong danh sách dưới đây, hãy trả về một đối tượng JSON chứa các trường sau:\n

1. `cpu_brand`: (String) Thương hiệu CPU ("Intel", "AMD", "Apple").\n
2. `cpu_series`: (String | null) Dòng CPU (ví dụ: "Core i5", "Ryzen 7", "M2", "M3", "M4", "Core Ultra 7"). Với Apple, nếu chỉ có "CPU Apple M2 8 nhân", series có thể là "M2" hoặc "null". Nếu chỉ có "CPU 10 lõi...", series có thể là `null`.\n
3. `cpu_model`: (String | null) Model cụ thể (ví dụ: "13420H", "7435HS", "1235U", "155H", "125H"). Đối với Apple, nếu không có model cụ thể ngoài series (như "M2", "M3", "M4"), trường này là `null`. Đối với "CPU 10 lõi...", model là `null`.\n
4. `cpu_cores`: (Integer | null) Tổng số lõi CPU. Trích xuất số lõi, ví dụ từ "8 lõi", "10 nhân", "12 cores", "4P+8E cores" (tính tổng 4+8=12).\n
5. `cpu_threads`: (Integer | null) Tổng số luồng CPU. Trích xuất từ "12 luồng", "16 luồng", "12T". Nếu không có, để `null`.\n
6. `cpu_base_clock`: (Float | null) Xung nhịp cơ bản tính bằng GHz. Trích xuất từ các giá trị như "3.1GHz", "2.6 GHz", "1.3 GHz", "P-core 1.4 / 5.0GHz" (lấy 1.4).\n
7. `cpu_boost_clock`: (Float | null) Xung nhịp tối đa/boost tính bằng GHz. Trích xuất từ "up to 4.60 GHz", "up to 4.5 GHz", "Max Turbo Frequency: 4.40 GHz", "P-core 1.4 / 5.0GHz" (lấy 5.0).\n
8. `raw_string`: (String) Chuỗi mô tả CPU gốc.\n

Yêu cầu quan trọng:\n

* Xác định chính xác `cpu_brand`.\n
* Với Apple CPU:\n
    * Nếu là "Apple M2 8 nhân", `cpu_brand` là "Apple", `cpu_series` là "M2", `cpu_model` có thể là `null`, `cpu_cores` là 8.\n

* Nếu là "CPU 10 lõi với 4 lõi hiệu năng và 6 lõi tiết kiệm điện", `cpu_brand` là "Apple", `cpu_series` có thể là "Apple Custom" hoặc `null`, `cpu_model` là `null`, `cpu_cores` là 10.\n

* Tính tổng số lõi nếu có cấu trúc P-core và E-core (ví dụ: "4P+8E lõi" -> `cpu_cores: 12`).\n

* Luôn trích xuất số lõi và số luồng dưới dạng số nguyên (Integer).\n

* Luôn trích xuất xung nhịp dưới dạng số thực (Float) và tính bằng GHz. Nếu có dải xung nhịp (ví dụ: "3.6GHz~4.9GHz"), `cpu_base_clock_ghz` là 3.6, `cpu_boost_clock_ghz` là 4.9. Nếu là "P-core 1.4 / 5.0GHz, E-core 0.9 / 3.8GHz", lấy giá trị P-core cho base và boost.\n

* Xử lý các đơn vị như "MHz" (hiếm) nếu có và chuyển sang GHz.\n

* Nếu thông tin không có hoặc không thể trích xuất, sử dụng `null`.\n

* Chỉ trả về một danh sách (list) các đối tượng JSON, mỗi đối tượng tương ứng với một chuỗi input. Không thêm giải thích nào khác.\n

"'

Cột ban đầu	Ví dụ	Các cột sau khi xử lý
cpu	AMD Ryzen 7 7435HS 3.1GHz (20MB Cache, up to 4.5 GHz, 8 lõi, 16 luồng)	cpu_brand cpu_type cpu_thread cpu_core cpu_boost_clock cpu_base_clock
laptop_cpu	AMD Ryzen 7	

3. **Rule-based**

Bảng thông số:

Dòng máy	cpu_thread	cpu_speed	cpu_max_speed
M1	8 (4P + 4E)	3.0	3.2
M1 pro	8 (6P + 2E) hoặc 10	3.0	3.2

	(8P + 2E)		
M1 max	10 (8P + 2E)	3.0	3.2
M1 ultra	20 (16P + 4E)	3.0	3.2
M2	8 (4P + 4E)	3.2	3.5
M2 pro	10 (6P + 4E) hoặc 12 (8P + 4E)	3.2	3.5
M2 max	12 (8P + 4E)	3.2	3.5
M2 ultra	24 (16P + 8E)	3.2	3.5
M3	8 (4P + 4E)	3.4	4.05
M3 pro	11 (5P + 6E) hoặc 12 (6P + 6E)	3.4	4.05
M3 max	14 (10P + 4E) hoặc 16 (12P + 4E)	3.4	4.05
M3 ultra	14 (10P + 4E) hoặc 16 (12P + 4E)	3.4	4.05
M4	10 (4P + 6E) hoặc 9 (3P + 6E)	3.5	4.4
M4 pro	14 (10P + 4E) hoặc 12 (8P + 4E)	3.8	4.5
M4 max	16 (12P + 4E) hoặc 14 (10P + 4E)	4.0	4.5

4. **Manual**

- Các trường hợp đặc biệt khi kiểm tra sau khi parsing

II. Data Ingestion

Data trước khi gộp:

Tên bộ dữ liệu	Số dòng	Số cột	Các thuộc tính
cellphones_laptop (1)	737	54	['**product_id**', '**name**', '**image**', 'display_size', 'storage_gb', 'key_selling_points', 'cam_ung', 'material', 'laptop_special_feature',

			'laptop_nganh_hoc', 'ram_storage', 'ram_max_support', 'product_state', 'storage_max_support', 'nhu_cau_su_dung', 'manufacturer', 'filter_label', 'cpu_brand', 'is_installment', 'filter_id', 'warranty_info', 'product_weight', **'url_path'**, 'url_key', 'os_version', 'display_width', 'vga_brand', 'bluetooth_version', 'has_bluetooth', 'display_height', 'ram_type', 'cpu_series', 'o_cung_laptop', 'vga_type', 'ports_slots', 'width_mm', 'height_mm', 'depth_mm', 'display_type', 'battery_capacity', 'laptop_camera', 'laptop_cong_nghe_am_thanh', 'ram_slots', 'cpu_model', 'ram_speed', 'cpu_cores', 'laptop_tam_nen_man_hinh', **'root_price'**, 'discounted_price', 'cpu_max_speed', 'refresh_rate', 'cpu_threads', 'cpu_speed', 'vga_vram']
cellphones_laptop_variant (2)	805	8	[**'root-laptop_id', 'child-laptop_id'**, **'child_laptop_name'**, **'child_laptop_image'**, **'child_laptop_link'**, 'child_laptop_color', **'child_laptop_price'**, **'special_features'**]
tgdd_laptop (3)	450	44	'filename', 'filename_origin', 'time_scraping', 'link', 'tieu_de', 'da_ban', 'gia_hien_tai', 'gia_cu', 'giam_gia', 'cong_nghe_cpu', 'so_nhan', 'so_luong', 'toc_do_cpu', 'toc_do_toi_da', 'ram', 'loai_ram', 'toc_do_bus_ram', 'ho_tro_ram_toi_da', 'o_cung', 'man_hinh', 'do_phan_giai', 'tan_so_quet', 'do_phu_mau', 'cong_nghe_man_hinh', 'card_man_hinh', 'cong_nghe_am_thanh', 'cong_giao_tiep', 'ket_noi_khong_day', 'webcam', 'tinh_nang_khac', 'den_ban_phim', 'kich_thuoc', 'chat_lieu',

			'thong_tin_pin', 'he_dieu_hanh', 'thoi_diem_ra_mat', 'chi_tiet', 'mau', 'url_anh', 'npu', 'hieu_nang_xu_ly_ai_tops', 'khe_doc_the_nho', 'man_hinh_cam_ung', 'tan_nhiet'

1. Gộp bảng (1) và (2) bằng left join trên product_id và root_laptop_id, tạo thành bảng cellphones_laptop (4)
   - Nếu tồn tại child_laptop_id, các cột root_id, name, image, url_path, root_price sẽ được lấy từ laptop_variant, sau đó xóa các cột thuộc bảng (2)
   - Đổi tên child_laptop_color thành laptop_color
   - Nếu child_laptop_id không tồn tại, gán mặc định laptop_color = 'đen'.

Tên bộ dữ liệu	Số dòng	Số cột	Các thuộc tính
cellphones_laptop (4)	904	54	['product_id', 'name', 'image', 'display_size', 'storage_gb', 'key_selling_points', 'cam_ung', 'material', 'laptop_special_feature', 'laptop_nganh_hoc', 'ram_storage', 'ram_max_support', 'product_state', 'storage_max_support', 'nhu_cau_su_dung', 'manufacturer', 'filter_label', 'cpu_brand', 'is_installment', 'filter_id', 'warranty_info', 'product_weight', 'url_path', 'url_key', 'os_version', 'display_width', 'vga_brand', 'bluetooth_version', 'has_bluetooth', 'display_height', 'ram_type', 'cpu_series', 'o_cung_laptop', 'vga_type', 'ports_slots', 'width_mm', 'height_mm', 'depth_mm', 'display_type', 'battery_capacity', 'laptop_camera', 'laptop_cong_nghe_am_thanh', 'ram_slots', 'cpu_model', 'ram_speed', 'cpu_cores', 'laptop_tam_nen_man_hinh', 'root_price', 'discounted_price', 'cpu_max_speed', 'refresh_rate', 'cpu_threads', 'cpu_speed',

			'vga_vram', ‘laptop_color’]
tgdd_laptop (3)	450	44	'product_id', 'time_scraping', 'url_path', 'name', 'da_ban', 'discounted_price', 'root_price', 'giam_gia', 'cpu_cores', 'cpu_threads', 'cpu_speed', 'cpu_max_speed', 'ram_storage', 'ram_speedm, 'max_ram_support', 'display_size', 'refresh_rate', 'do_phu_mau', 'display_type','laptop_cong_nghe_am_thanh', 'ports_slots', 'laptop_special_feature', 'den_ban_phim', 'material', 'thoi_diem_ra_mat', 'key_selling_points', 'laptop_color', 'image', 'npu', 'khe_doc_the_nho', 'cam_ung', 'tan_nhiet', 'cpu_brand', 'cpu_series', 'cpu_architecture', 'cpu_model', 'hieu_nang_xu_ly_ai_tops', 'ram_type', 'ram_slots', 'drive_type', 'interface', 'storage_gb', 'storage_max_support', 'display_width', 'display_height', 'color_gamut_score', 'vga_type', 'vga_brand', 'vga_vram', 'wifi_version', 'bluetooth_version', 'width_mm', 'height_mm', 'depth_mm', 'product_weight', 'os_version', 'battery_capacity', 'laptop_camera', 'product_state', 'manufacturer', 'is_installment', 'nhu_cau_su_dung'

2. Gộp (3) và (4) → laptop_bronze: Chỉ giữ lại các cột ở cả 2 file cùng có, loại bỏ các cột thừa

Tên file	Số dòng	Số cột	Các thuộc tính
laptop_bronze	1353	40	['cam_ung', 'height_mm', 'url_path', 'material', 'storage_max_support', 'storage_gb', 'display_width', 'bluetooth_version', 'width_mm', 'manufacturer', 'cpu_max_speed', 'root_price', 'cpu_threads', 'cpu_cores', 'ram_speed', 'product_weight', 'discounted_price', 'ram_type', 'refresh_rate', 'cpu_speed', 'name', 'ram_storage',

			'is_installment', 'ram_slots', 'os_version', 'battery_capacity', 'laptop_color', 'cpu_model', 'nhu_cau_su_dung', 'vga_type', 'depth_mm', 'image', 'vga_brand', 'product_id', 'cpu_series', 'display_height', 'vga_vram', 'laptop_camera', 'display_size', 'cpu_brand']

III. Data Cleaning

1. Xử lý giá trị ngoại lai

- Phương pháp: xử lý thủ công
- Số cột có giá trị ngoại lai: 9 cột (các cột numerical continuous: 'cpu_max_speed', 'root_price', "height_mm', 'width_mm', 'discounted_price', 'battery_capacity', 'display_size', 'display_height', 'display_width'
- Số lượng giá trị ngoại lại: ~ < 10%/ cột
- Nguyên nhân:

+ Lỗi trong quá trình phân tách (parsing) dữ liệu
+ Đơn vị chưa được chuẩn hóa chính xác
+ Dữ liệu sai từ nguồn gốc

- Cách khắc phục: Tiến hành tra cứu thủ công trên các website chính hãng và chỉnh sửa lại dữ

2. Xử lý dữ liệu thiếu

a. Xử lý bằng LLMs

- Mô hình sử dụng: Gemini 2.0, Gemini 2.5, Llama 3.3
- Phương pháp: Few-shot Prompting, Chain-of-thought Prompting

prompt = f"""

You are a highly proficient AI assistant specialized in accurately retrieving and standardizing laptop technical specifications from web sources

Your task is to find missing technical specifications for a given laptop and return these specifications in a structured JSON format, adhering to strict data type and unit conversion rules

CONTEXT:

You are provided with data for a laptop entry that has some missing (NULL) specifications

Known information about this laptop includes:

- Laptop name: {laptop_name}
- Manufacturer: {manufacturer}
- Other known specifications: {existing_specs_str}

The following columns have missing information and require your lookup: {', '.join(null_columns_list)}

INSTRUCTIONS:

1. Web search: Based on the 'Laptop name', the 'Manufacturer' and any 'Other known specifications' (if needed), perform a targeted web search to find the values for the missing columns: {', '.join(null_columns_list)}

2. JSON output requirement:

- You MUST return the results as a single, valid JSON object
- The 'key' in the JSON object must be exactly match the column names provided in the missing columns list '{null_columns_list[0]}'
- If, after a thorough search, you cannot find reliable information for a specific column, use the JSON value 'null' (NOT the string "null") for that key

3. Data formatting and unit conversion (CRITICAL)

- For all numerical specifications, you MUST return ONLY THE NUMERICAL VALUE (as a float or integer). DO NOT include units (like "GB", "Hz", "mm", "GHz", "VND", "Wh") in the JSON value itself
- Adhere to the following specific instructions for each column type if it appears in the 'null_columns_list':

+ 'laptop_height_mm': report height in milimeters(mm). Return ONLY the numerical value as a float(for 23.5 cm , convert and turn '235.0')

+ 'laptop_width_mm': report width in milimeters(mm). Return ONLY the numerical value as a float(for 39.5 cm , convert and turn '395.0')

+ 'laptop_depth_mm': report depth/thickness in milimeters(mm). Return ONLY the numerical value as a float(for 1.75 cm , convert and turn '17.5')

+ 'cpu_boost_clock_ghz': Report CPU boost clock speed in Gigahertz (GHz). Return ONLY the numerical value as a float (e.g., for 4.9 GHz, return 4.9; for 5000 MHz, return 5.0)

+ 'root_price_vnd': Report original price in Vietnamese Dong (VND). Find the original price in USD, then you MUST convert into VND using the fixed conversion rate: 1 USD = 26000 VND. Return ONLY the numerical value as a float or integer, without commas or currency symbols (e.g., for $750 USD, return 19500000.0). Do not include any units or text.

+ 'cpu_threads': Report the number of CPU threads. Return ONLY the numerical value as an integer (e.g., 16)

+ 'cpu_cores': Report the number of CPU cores. Return ONLY the numerical value as an integer (e.g., 10)

+ 'ram_speed': Report RAM speed in Megahertz (MHz). Return ONLY the numerical value as a float or integer (for 4800 MHz, return 4800.0)

+ 'ram_type': (e.g., "ddr4", "lpddr5", "ddr5") Return as a lowercase string without spaces

+ 'refresh_rate_hz': Report display refresh rate in Hertz (Hz). Return ONLY the numerical value as an integer or float (e.g., for 60Hz, return 60.0; for 144Hz, return 144.0)

+ 'cpu_base_clock_ghz': Report CPU base clock speed in Gigahertz (GHz). Return ONLY the numerical value as a float (e.g., for 2.4 GHz, return 2.4; for 3200 MHz, return 3.2).

+ 'ram_slots': Report the number of RAM slots. Return ONLY the numerical value as an integer (e.g., 0, 1, 2). If RAM is soldered and no slots, return 0.

+ 'battery_capacity_wh': Report battery capacity in Watt-hours (Wh). Return ONLY the numerical value as a float (e.g., for 47Wh, return 47.0)

+ 'cpu_model': (e.g., "13620h", "i7-12700h", "ryzen 7 5800u") Return as a string, preferably lowercase and normalized (e.g., remove "Intel Core", "AMD Ryzen" prefixes if the manufacturer column already specifies this, focus on the model number/name).

+ 'vga_type' This field MUST be one of two specific string values:

- If the laptop uses integrated graphics (e.g., Intel Iris Xe, AMD Radeon Graphics), return "card tích hợp"
- If the laptop has a dedicated/discrete graphics card (e.g., NVIDIA GeForce RTX 3050, AMD Radeon RX 6600M), return "card rời"
- Do not use any other values.

+ 'vga_vram_gb': Report dedicated Video RAM (VRAM) capacity in Gigabytes (GB). Return ONLY the numerical value as a float (e.g., for 4GB VRAM, return 4.0). If vga_type is "card tích hợp", this value should typically be 0.0 as integrated graphics share system memory. Clarify if "shared memory" should be reported here or if it should be strictly for dedicated VRAM. For now, assume dedicated VRAM; if none, return null

+ 'laptop_camera': This field MUST be one of two specific string values:

- If the camera is Full HD (typically 1080p), return "full hd".
- If the camera is HD (typically 720p), return "hd".
- If other (e.g., VGA, or specific resolution like 5MP), and no direct match to "hd" or "full hd", attempt to classify or return the specific resolution as a string if these two are not applicable. However, strongly prefer "hd" or "full hd". If uncertain between the two, or if it's a higher non-standard resolution, you may return the most descriptive common term.

+ 'discounted_price_vnd': Report discounted price in Vietnamese Dong (VND). Find the discounted price in USD, then you MUST convert into VND using the fixed conversion rate: 1 USD = 26000 VND. Return ONLY the numerical value as a float or integer, without commas or currency symbols (e.g., for $750 USD, return 19500000.0). Do not include any units or text.

4. Chain of thought:

Step 01: Understand the data requirements and the format rules for each missing field

Step 02: Perform a web search (oor rely on your interrnal knowledge) using the provided laptop name, manufacturer, and known specification

Step 03: Summerize and extract relevant technical specifications based on your findings

Step 04: The price is found in USD, convert it into VND using the fixed conversion rate: 1 USD = 26000VND

Step 04: Format and return the completed information as a valid JSON object, following all formatting and unit conversion instructions precisely

5. Data source priority: Proritize information form official manufacturer websites, reputable e-commerce listings (from the manufacturer or major retailers), or well-known tech reiview sites

INPUT DATA EXAMPLE (Just illustrative, showing how information might be structured for a request):

Laptop name: Lenovo Legion 5 Pro 16ACH6H

Manufacturer: Lenovo

Known Specs:

- cpu_model: i5-12500H
- ram_type: ddr5

Null Columns to fill: ["ram_speed", "vga_type", "vga_vram_gb", "refresh_rate_hz", "laptop_camera", "battery_capacity_wh"]

EXPECTED JSON OUTPUT FORMAT (This is the precise format you MUST return):

`json

{{

"ram_speed": 4800.0,

"vga_type": "card rời",

"vga_vram_gb": 6.0,

"refresh_rate_hz": 144.0,

"laptop_camera": "hd",

"battery_capacity_wh": 57.5

}}

"""

Kết quả: Độ chính xác được đánh giá bằng cách đối chiếu kết quả với 30% dữ liệu đã được xử lý thủ công

Gemini 2.5	98,42%
Gemini 2.0	97,60%
Llama 3.3	62,90%

b. Xử lý bằng thuật toán

Điền giá trị thiếu theo nhóm đặc trưng

Tên cột	**Nhóm thuộc tính phụ thuộc**	**Phương pháp**
vga_type	cpu_model, vga_brand, manufacturer	rule-based
laptop_camera	manufacturer	mode
width_mm, height_mm	display_size	mean
depth_mm	nhu_cau_su_dung (mong_nhe, gaming)	rule-based
cpu_model	name, manufacturer	rule-based, manual
ram_type	cpu_model	rule-based
ram_slots	name, display_size	rule-based
Cpu_cores, cpu_threads	cpu_series, ram_storage, cpu_model, cpu_brand, nhu_cau_su_dung, display_size	machine learning (XGBoost)
cpu_speed, cpu_max_speed	cpu_cores, cpu_model, cpu_series	rule-based. machine learning (XGBoost)
refresh_rate	vga_brand, cpu_series,display_size	rule-based, mode
battery_capacity	vga_type, nhu_cau_su_dung, product_weight	median

ram_speed	ram_type	median
vga_vram	vga_type, cpu_model, cpu_series, ram_storage, vga_brand, nhu_cau_su_dung	
discounted_price, root_price	discounted_price, root_price	mean, LLMs
nhu_cau_su_dung	trừ các cột: url_path, product_id, image, name, nhu_cau_su_dung	embeddings (PhoBERT)
ram_storage, bluetooth_version, wifi_version		manual

3. Chuẩn hóa dữ liệu

Column name	Chuyển các tên thành kí tự tiếng việt không dấu, lower case, phân cách bằng "_" (tiếng anh - giữ nguyên) bỏ các ký tự /n trong tên cột
Price	Chuyển thành kiểu int, bỏ ký tự tiền tệ
Unit	- Dung lượng (RAM, ổ cứng): Chuẩn hóa về GB (ví dụ: 512 GB SSD, 16 GB RAM). Nếu có đơn vị TB, chuyển đổi thành GB (1 TB = 1024 GB). - Tần số (Tốc độ CPU, Bus RAM): Chuẩn hóa về GHz, đảm bảo đồng nhất. - Kích thước màn hình: Chuẩn hóa về inch (ví dụ: 14"" → 14) - Số lượng đã bán (Đã bán): Chuyển đổi 1,3k thành 1300 (số nguyên). - Giảm giá (Giảm giá (%)): Loại bỏ ký hiệu % và chuyển thành số thập phân (float), số dương - Bỏ các unit để phù hợp với mô hình Machine Learning - các cột binary thì chuyển thành 0, 1
Text	Xử lý giá trị text: Chuyển tất cả text về dạng viết thường (lowercase) để tránh phân biệt hoa thường (ví dụ: Intel Core i5 → intel core i5) - Không có thông tin thì chuyển thành : Hãng không công bố

Time	Xử lý thời gian: Cột time_scraping chuyển về định dạng datetime (YYYY-MM-DD HH:MM:SS)

IV. Data Preprocessing

- One-hot encoding theo nhu cầu để phân tích kỹ lưỡng sự ảnh hưởng với biến target:

Cột ban đầu	Cột sau khi xử lý
nhu_cau_su_dung	mong_nhe, do_hoa_ky_thuat, gaming, hoc_tap_van_phong, cao_cap_sang_trong, laptop_sang_tao_noi_dung

- **Feature Selection**:.
  + Thực hiện kiểm định ANOVA với các categorical features với giả thiết với biến target "root_price":
    H0: hai biến độc lập
    H1: hai biến phụ thuộc nhau

Index	Features	F-statistic	P-value	Đánh giá
0	material	92,1582	0	Ảnh hưởng
1	manufacturer	19,3462	0	Ảnh hưởng
2	ram_type	46,2786	0	Ảnh hưởng
3	os_version	52,236	0	Ảnh hưởng
4	laptop_color	2,1626	0,018	Ảnh hưởng
5	vga_type	0,017	0,8962	Không ảnh hưởng
6	vga_brand	66,6873	0	Ảnh hưởng
7	laptop_camera	247,5311	0	Ảnh hưởng
8	cpu_brand	67,3149	0	Ảnh hưởng

- Feature importance cho biến root_price

Features	Feature importances
ram_storage	0,544631
display_width	0,116766
display_height	0,054055
storage_max_support	0,045027
storage_gb	0,042426
cpu_cores	0,036955
battery_capacity	0,028198
product_weight	0,017607
cpu_threads	0,01035

- **Kết quả:**
  + Các biến được chọn làm feature (input):

    'storage_max_support', 'storage_gb', 'display_width', 'cpu_threads', 'cpu_cores', 'ram_speed', 'cpu_speed', 'ram_storage', 'ram_slots', 'battery_capacity', 'display_height', 'laptop_sang_tao_noi_dung', 'do_hoa_ky_thuat', 'cao_cap_sang_trong', 'material', 'manufacturer', 'ram_type', 'os_version', 'laptop_color', 'vga_brand', 'laptop_camera', 'cpu_brand'

  + Biến target (output): ‘root_price’
- **Encoding**: One-Hot Encoding cho biến phân loại

'manufacturer', 'cpu_brand', 'material', 'os_version', 'laptop_color', 'vga_brand', 'laptop_camera', 'ram_type', 'laptop_sang_tao_noi_dung', 'do_hoa_ky_thuat', 'cao_cap_sang_trong'

- **Scaling**: StandardScaler cho biến số

  'storage_max_support', 'storage_gb', 'display_width', 'cpu_threads', 'cpu_cores', 'ram_speed', 'cpu_speed', 'ram_storage', 'ram_slots', 'battery_capacity', 'display_height'