Tiểu luận cuối kỳ

Học phần: Khoa học dữ liệu

Tên đề tài: Prediction of laptop’s price

Nhóm 02

1. **Nguyễn Văn Mạnh**
2. **Nguyễn Công Cường**
3. **Phan Tiến Đạt**

2023

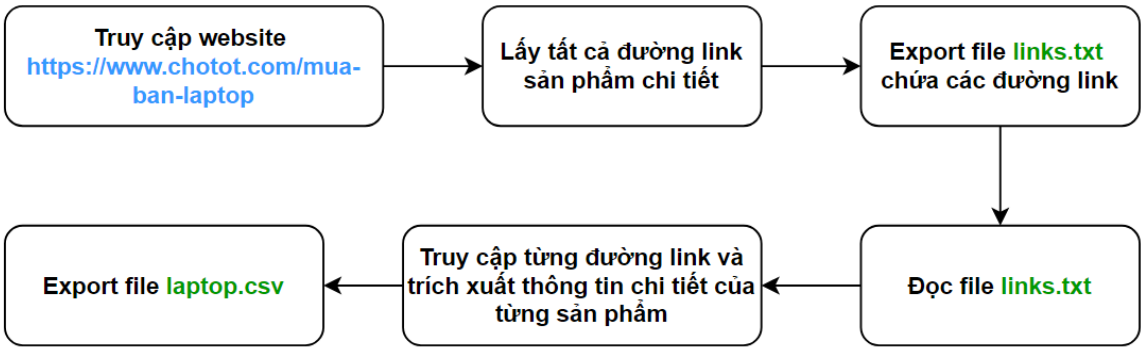
BẢNG PHÂN CÔNG NHIỆM VỤ

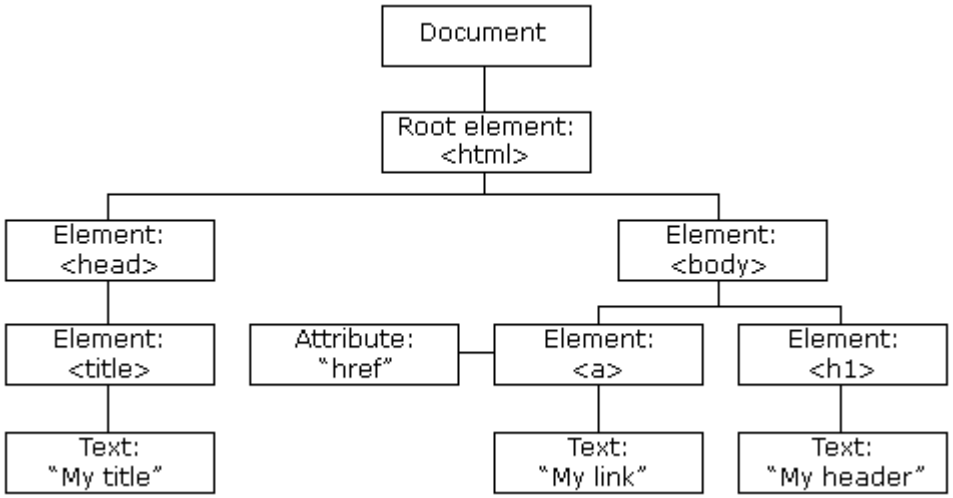
STT	Sinh viên thực hiện	Các nhiệm vụ	Tự đánh giá
01	Nguyễn Văn Mạnh	Cào dữ liệu từ trang website Làm sạch dữ liệu thô Trực quan hoá dữ liệu	Đã hoàn thành
02	Phan Tiến Đạt	Làm sạch và xử lý dữ liệu trống Xử lý ngoại lệ Chuẩn hóa Lựa chọn thuộc tính Giảm chiều dữ liệu Thể hiện hiệu quả của các quá trình tiền xử lý	Đã hoàn thành
03	Nguyễn Công Cường	Khảo sát mô hình. Cài đặt mô hình. Lựa chọn mô hình, điều chỉnh siêu tham số sử dụng GridSearchCV. Trực quan hóa kết quả dự đoán trên các mô hình. Tìm hiểu và tính toán các metrics để đưa ra so sánh, nhận xét.	Đã hoàn thành

Prediction of laptop's price

1. Thu thập và mô tả dữ liệu

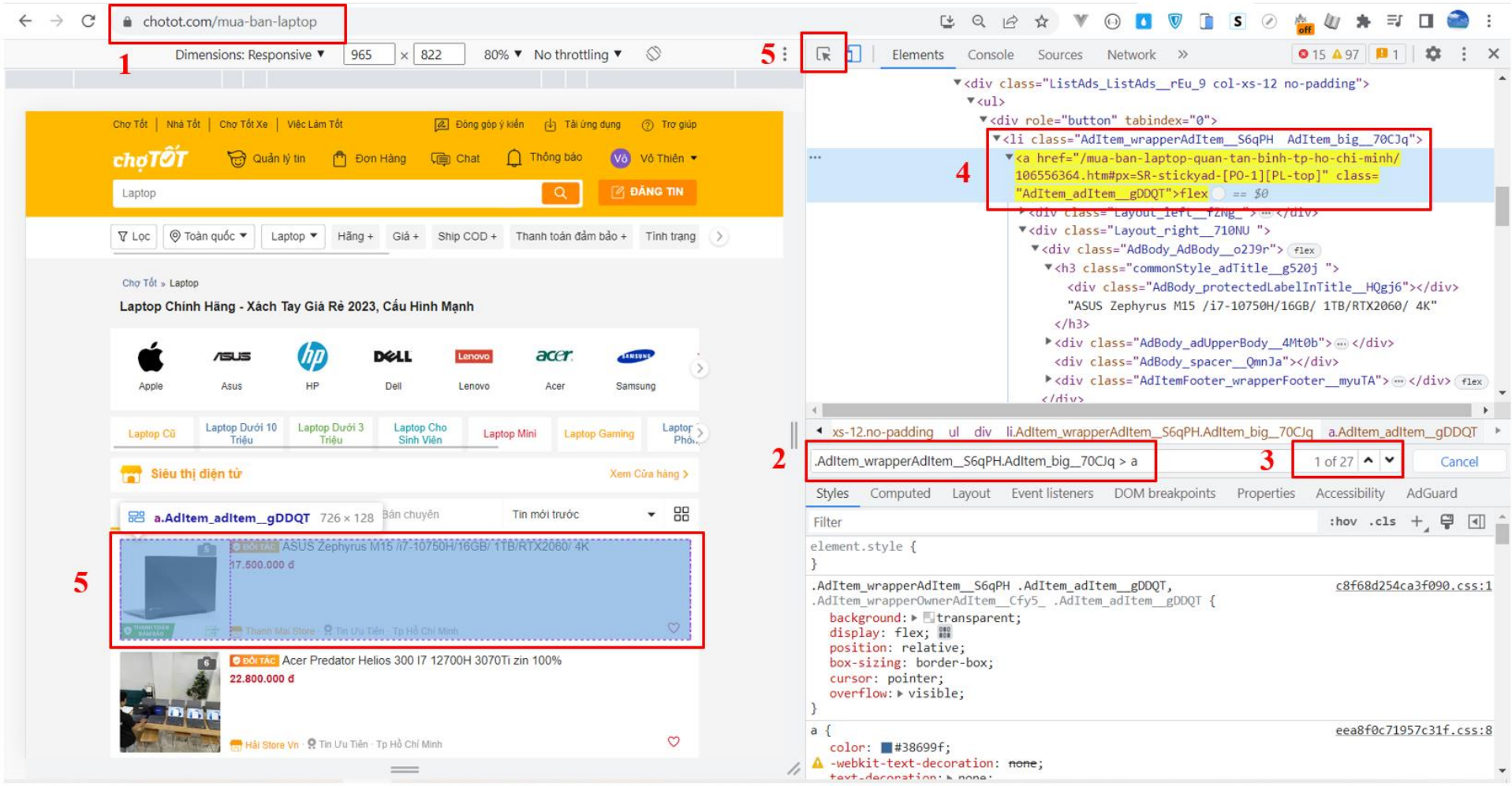
- **Các bước thập dữ liệu**





Prediction of laptop's price

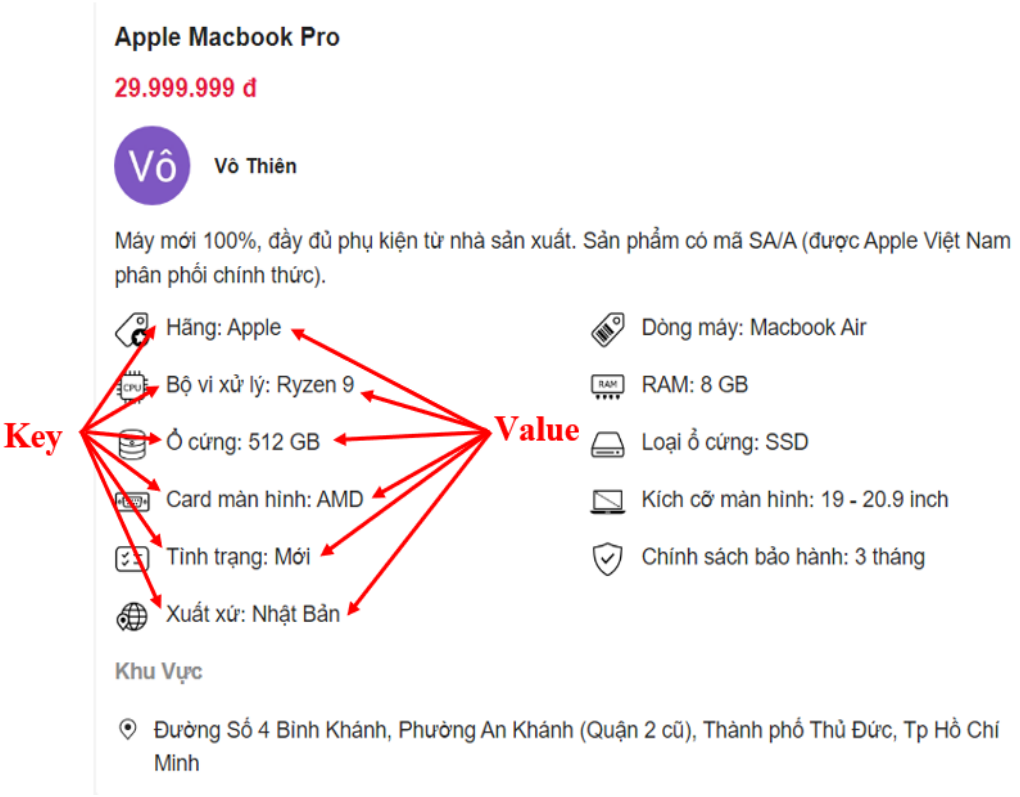
Bước 1: Thu thập tất cả các đường link sản phẩm

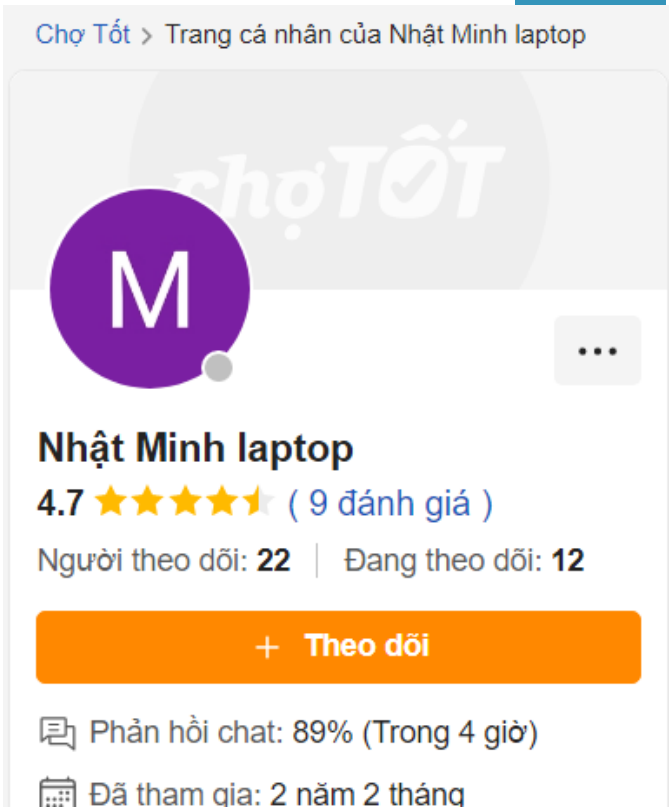


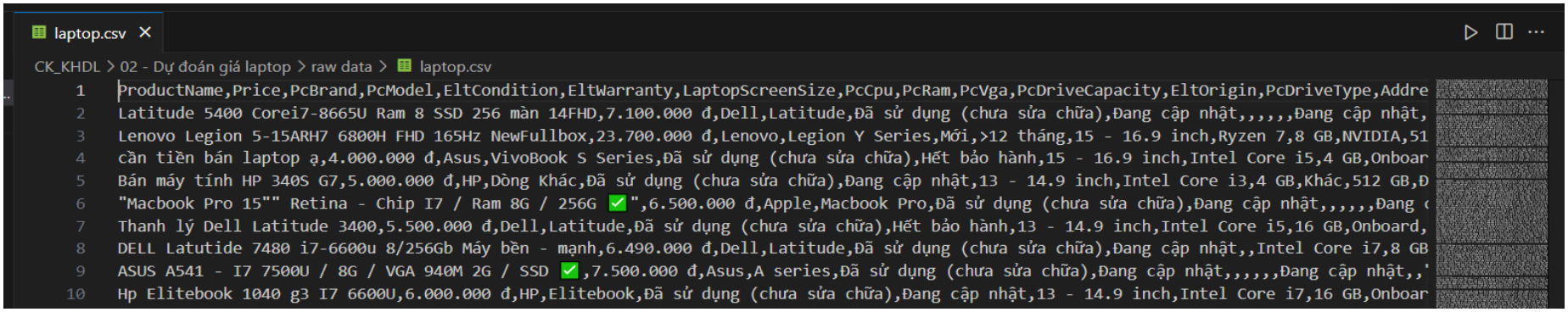
Tiếp đến tiến hành sử dụng **Selenium** để giả lập người dùng duyệt website, và sử dụng các phương thức như **find_elements(By.CSS_SELECTOR,"[class= 'AdItem_wrapperAdItem__S6qPH AdItem_big__70CJq'] > a")** để bóc tách đường link.

Prediction of laptop's price

Bước 2: Lấy thông tin chi tiết sản phẩm







Prediction of laptop's price

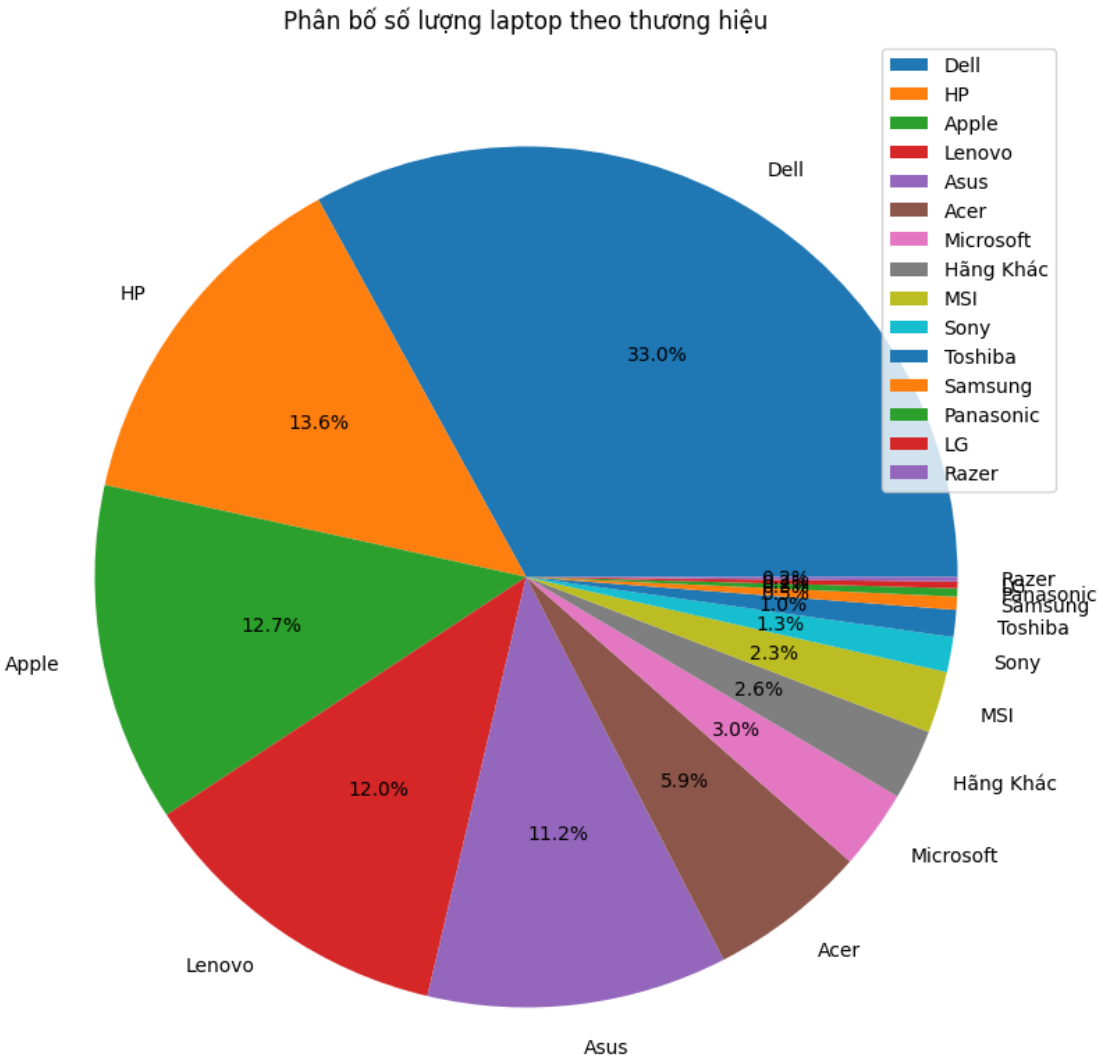
Mô tả dữ liệu

- Số lượng đặc trưng: **15**
- Số lượng mẫu: **12128**

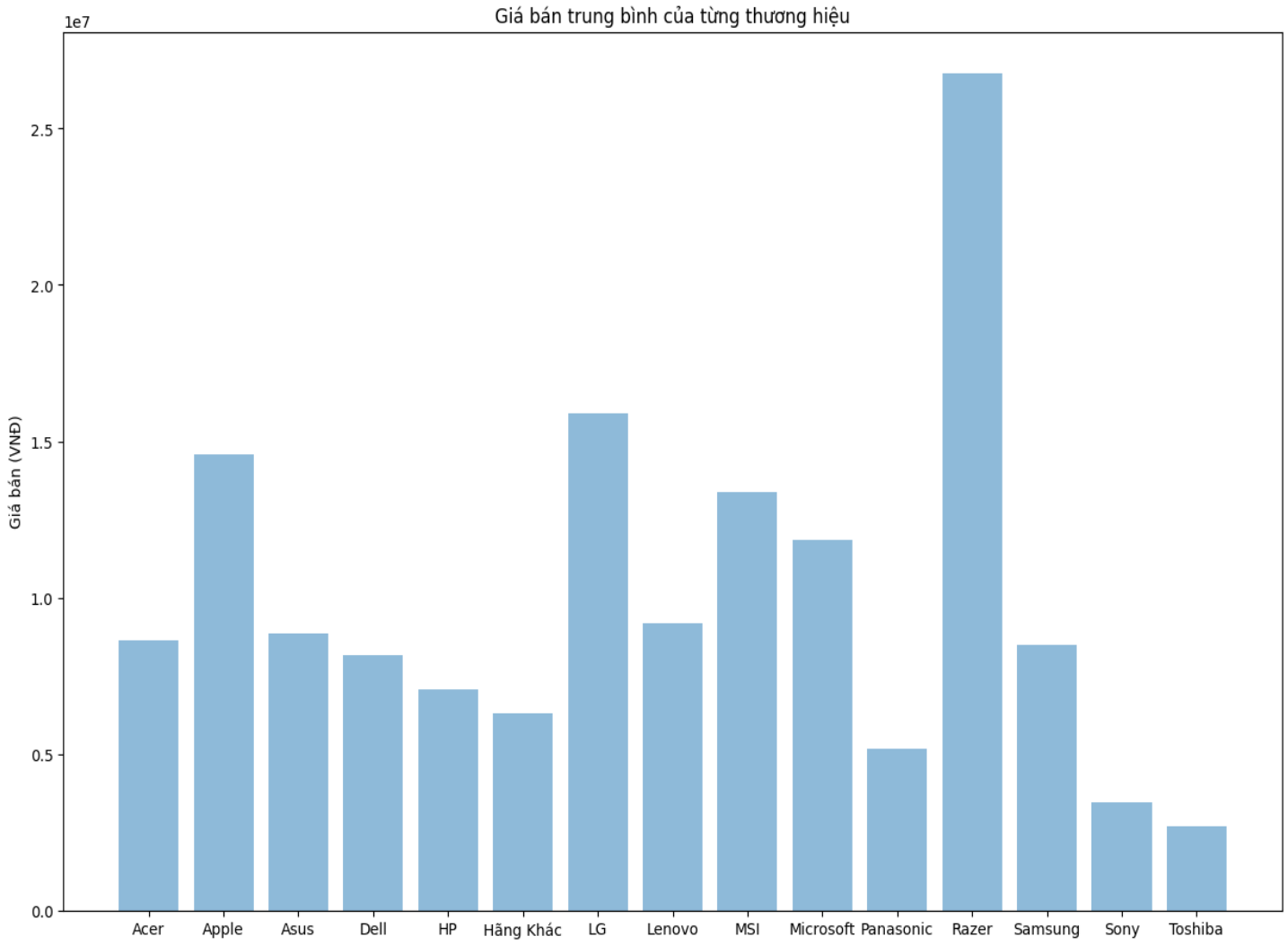
STT	Đặc trưng	Mô tả	Kiểu dữ liệu	Số mẫu dl trống
1	ProductName	Tên laptop	String	0
2	Price	Giá bán (vnđ)	Float	0
3	PcBrand	Thương hiệu laptop	String	0
4	PcModel	Mẫu laptop	String	2
5	EltCondition	Trạng thái laptop	String	0
6	EltWarranty	Thời gian bảo hành	Float	3265
7	LaptopScreenSize	Kích thước màn hình	Float	268
8	PcCpu	Tên chip xử lý	String	91
9	PcRam	Dung lượng RAM (GB)	Float	38
10	PcVga	Tên Card màn hình	String	977
11	PcDriveCapacity	Dung lượng ổ cứng	Float	87
12	EltOrigin	Xuất xứ	String	11334
13	Address	Địa chỉ thành phố	String	0
14	ShopRating	Số sao đánh giá	Float	3059
15	Comments	Số bình luận đánh giá	Float	3262

Prediction of laptop's price

- **Trực quan hóa dữ liệu**



Prediction of laptop's price



Prediction of laptop's price

- **2. Trích xuất đặc trưng**
  - **Dữ liệu trống tồn tại trong dataset**
    - BigDS

```
Price                0
PcBrand              0
PcModel              1
EltCondition         0
EltWarranty       2755
LaptopScreenSize   177
PcCpu               70
PcRam               32
PcVga              707
PcDriveCapacity     64
EltOrigin         9299
Address              0
ShopRating        2012
Comments          2174
dtype: int64
```

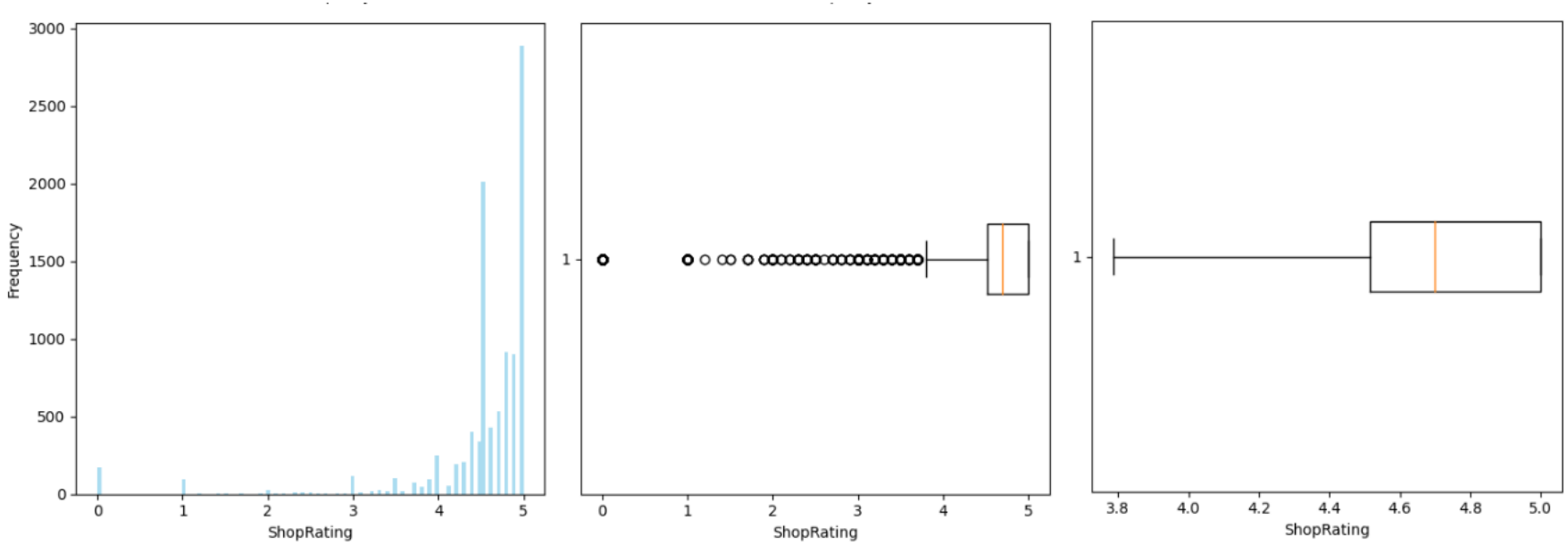
    - SmallDS

```
Price                0
PcBrand              0
PcModel              0
EltCondition         0
EltWarranty        250
LaptopScreenSize    25
PcCpu               10
PcRam                3
PcVga               65
PcDriveCapacity      3
EltOrigin          919
Address              0
ShopRating         206
Comments           227
dtype: int64
```

  - **Xử lý dữ liệu trống bằng 7 kĩ thuật khác nhau**
    - Mean
    - Median
    - Any
    - Mode
    - Random
    - Arbitrary
    - Iterative

Prediction of laptop's price

- **Xử lí ngoại lệ**
  - Phân bố của ShopRating và Boxplot trước khi xử lý ngoại lệ
  - Sau khi xử lý ngoại lệ



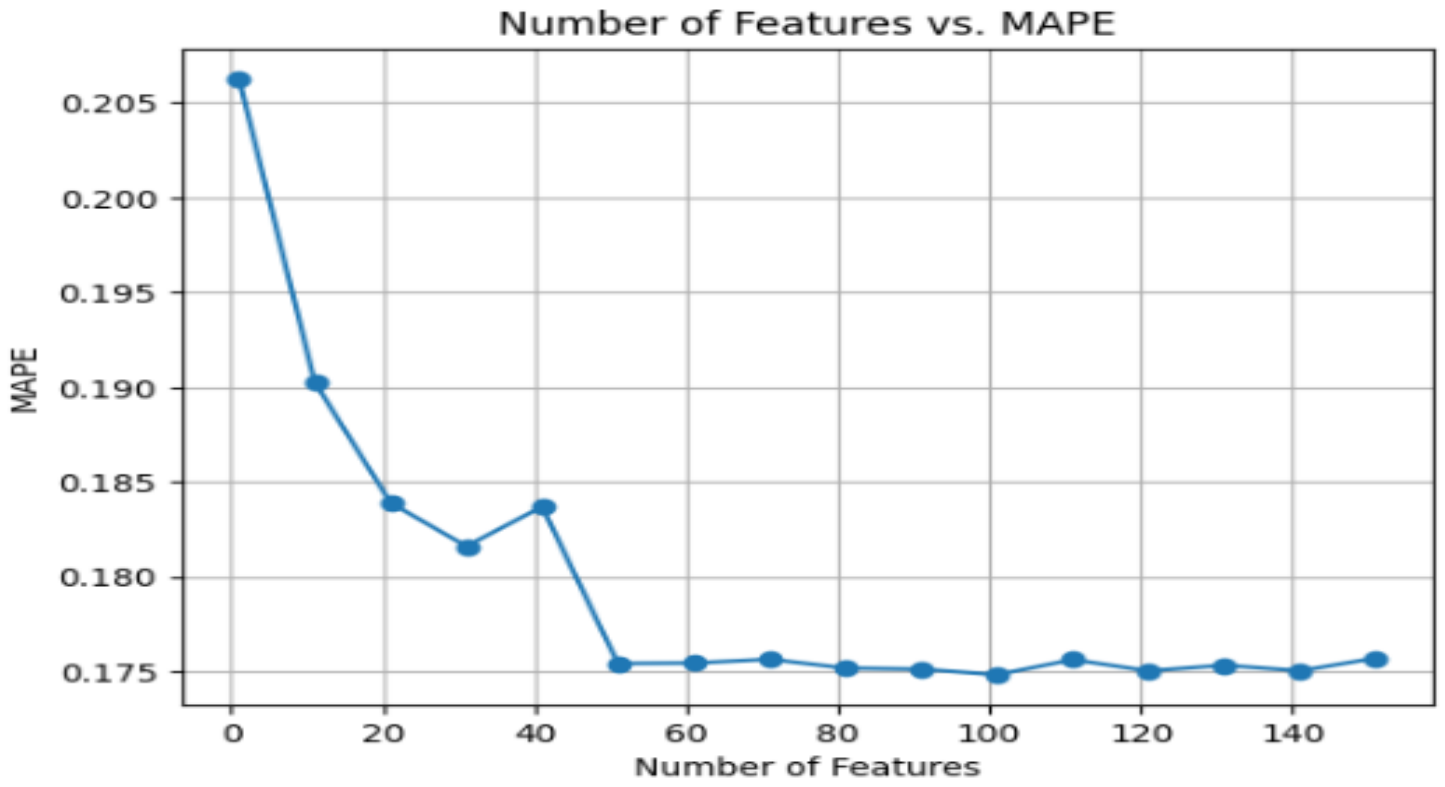
Prediction of laptop's price

- **Các loại chuẩn hoá được áp dụng**
  - StandardScaler, MinMaxScaler, RobustScaler, MaxAbsScaler, Normalizer, QuantileTransformer và PowerTransformer
  - Đánh giá tổ hợp các phương pháp sử dụng MAPE

	mean	median	mode	any	random	arbitrary	iterative
StandardScaler	0.175019	0.179001	0.179422	0.176732	0.177459	0.175757	0.175019
MinMaxScaler	0.174996	0.179086	0.179395	0.176331	0.177486	0.175714	0.174996
RobustScaler	0.175007	0.179011	0.179282	0.176400	0.177502	0.175585	0.175007
MaxAbsScaler	0.174963	0.179024	0.179433	0.176557	0.177519	0.175701	0.174963
Normalizer	0.177867	0.181483	0.182122	0.184369	0.183972	0.178235	0.177867
QuantileTransformer	0.174980	0.178835	0.179526	0.175990	0.177354	0.175644	0.174980
PowerTransformer	0.174943	0.179043	0.179276	0.176103	0.177729	0.175711	0.174943

Prediction of laptop's price

- **Lựa chọn đặc trưng**
  - Phương pháp được sử dụng là SelectKBest và RFE (Recursive Feature Elimination)
  - Độ thị độ lỗi MAPE theo số đặc trưng được chọn



Prediction of laptop's price

- **Giảm chiều dữ liệu**
  - Phương pháp được sử dụng: Principal component analysis (PCA)
  - PCA không đem lại hiệu quả trong đề tài

```
PCA 1 : độ lỗi là 0.965495255288441 cải thiện -0.7906562808251056
PCA 21 : độ lỗi là 0.5048669647364341 cải thiện -0.33002799027309865
PCA 41 : độ lỗi là 0.40693823781221394 cải thiện -0.23209926334887848
PCA 61 : độ lỗi là 0.399232452999869 cải thiện -0.22439347853653352
PCA 81 : độ lỗi là 0.3957667649395962 cải thiện -0.22092779047626077
```

```
PCA 1 : độ lỗi là 0.6024928712468125 cải thiện -0.4539189431425624
PCA 11 : độ lỗi là 0.5995589153239442 cải thiện -0.45098498721969416
PCA 21 : độ lỗi là 0.5470124643171436 cải thiện -0.3984385362128935
PCA 31 : độ lỗi là 0.5460914329008166 cải thiện -0.39751750479656656
PCA 41 : độ lỗi là 0.5353811626090156 cải thiện -0.3868072345047655
PCA 51 : độ lỗi là 0.540342612601335 cải thiện -0.3917686844970849
PCA 61 : độ lỗi là 0.5497112671059537 cải thiện -0.4011373390017037
PCA 71 : độ lỗi là 0.524417805059933 cải thiện -0.3758438769556829
PCA 81 : độ lỗi là 0.5268129344577466 cải thiện -0.3782390063534965
PCA 91 : độ lỗi là 0.47011207736899374 cải thiện -0.32153814926474367
PCA 101 : độ lỗi là 0.4131018287371292 cải thiện -0.2645279006328791
```

Prediction of laptop's price

- **3. Mô hình hóa dữ liệu**
  - **Chia tập dữ liệu**
  - Tổng toàn bộ dữ liệu sau bước Feature engineering:

+ Big Data ~10000 mẫu, Small Data ~1000 mẫu

- Training set: 70%

+ Big Data ~7000 mẫu, Small Data ~700 mẫu

- Testing set: 30%

+ Big Data ~3000 mẫu, Small Data ~300 mẫu

- Training Validation set:

+ Big Data ~4900 mẫu, Small Data ~490 mẫu

- Test Validation set:

+ Big Data ~2100 mẫu, Small Data ~210 mẫu

Mô hình sử dụng và bộ tham số

- **Linear Regression:** Không có siêu tham số để điều chỉnh.
- **Random Forest Regressor:**

- n_estimators: Số lượng cây trong rừng ngẫu nhiên. Giá trị này thường được chọn lớn để đảm bảo tính ổn định của dự đoán. Ví dụ: [100, 200].

- max_depth: Độ sâu tối đa của cây. Giới hạn độ sâu này giúp tránh việc mô hình quá phức tạp và overfitting. Ví dụ: [None, 5, 10].

- min_samples_split: Số lượng mẫu tối thiểu cần có trong mỗi nút để tiếp tục quá trình chia.

Giá trị này có thể giúp kiểm soát độ phức tạp của cây.Ví dụ: [None, 5, 10].

- min_samples_leaf: Số lượng mẫu tối thiểu cần có trong mỗi lá để xem xét một phân vùng là một lá.

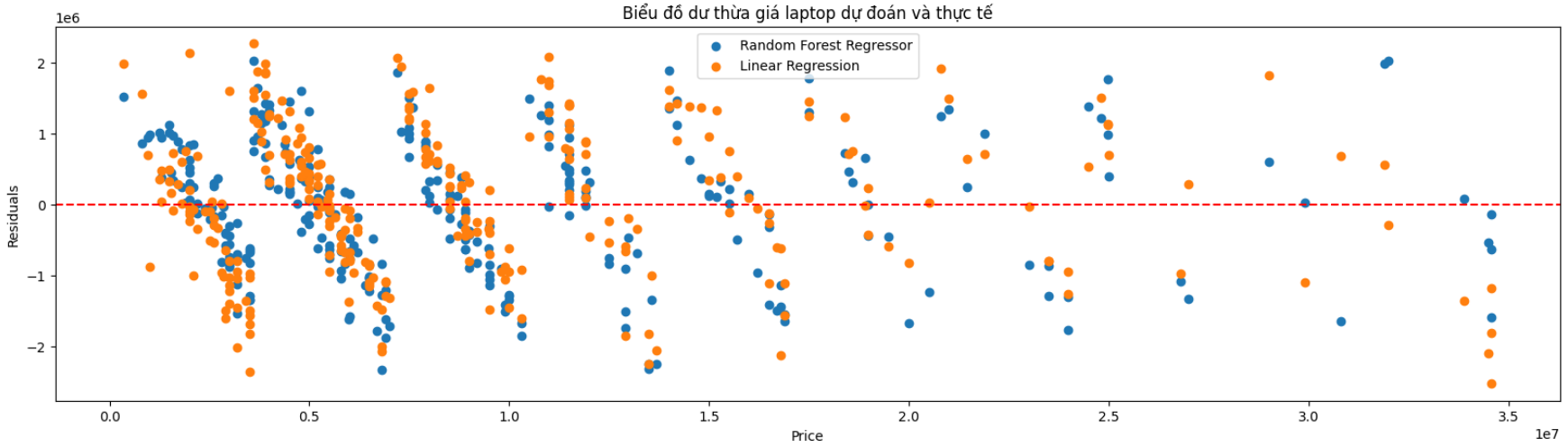
Giá trị này có thể giúp tránh việc quá khớp Ví dụ: [None, 5, 10].

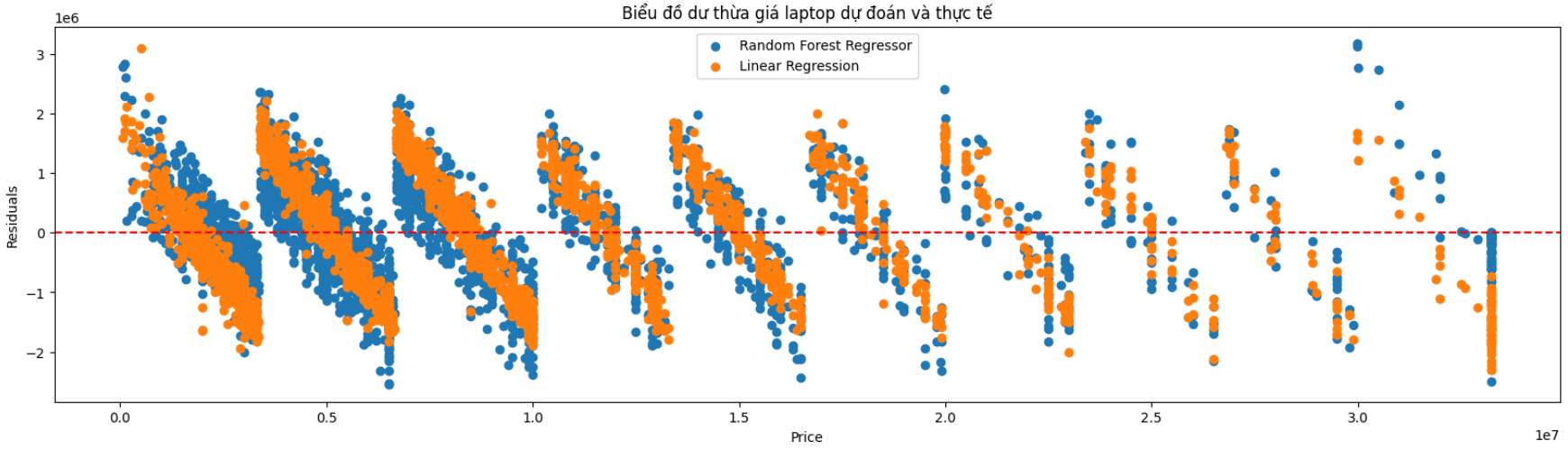
Prediction of laptop's price

- **Sử dụng GridSearchCV thu được bộ tham số tốt nhất:**
- **Small Data:**
- Best Hyperparameters:
- *{'max_depth': None,'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}*
- **Big Data:**
- Best Hyperparameters:
- *{'max_depth': None, 'n_estimators': 50}*
- Sau khi đã tìm ra bộ siêu tham số tốt nhất thì áp dụng để huấn luyện mô hình và dự đoán kết quả trên tập test

Prediction of laptop's price

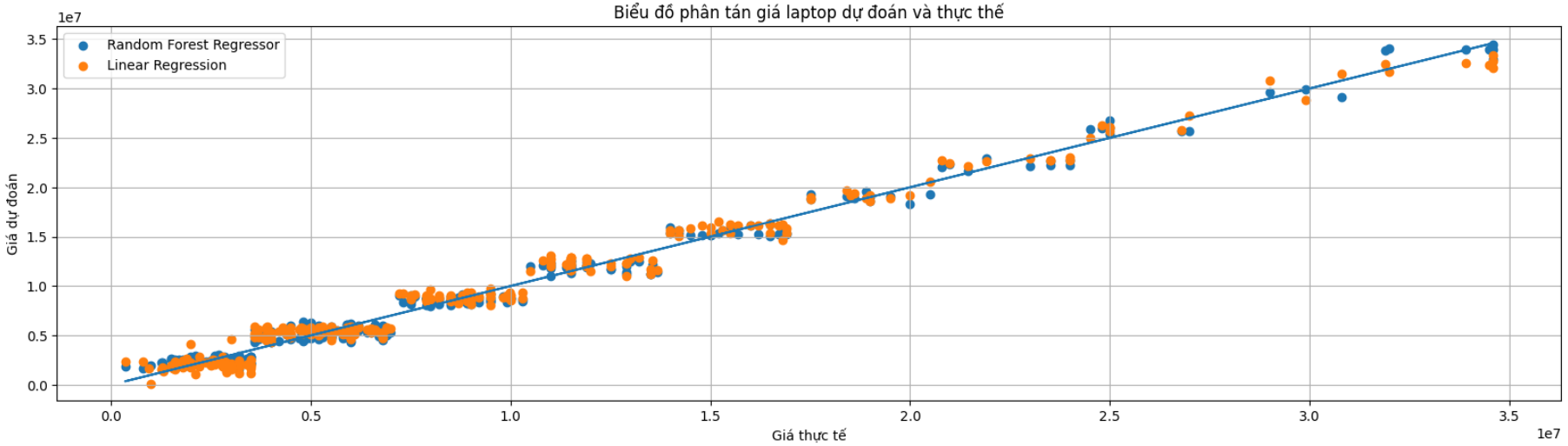
- **Đồ thị kết quả và so sánh 2 mô hình**

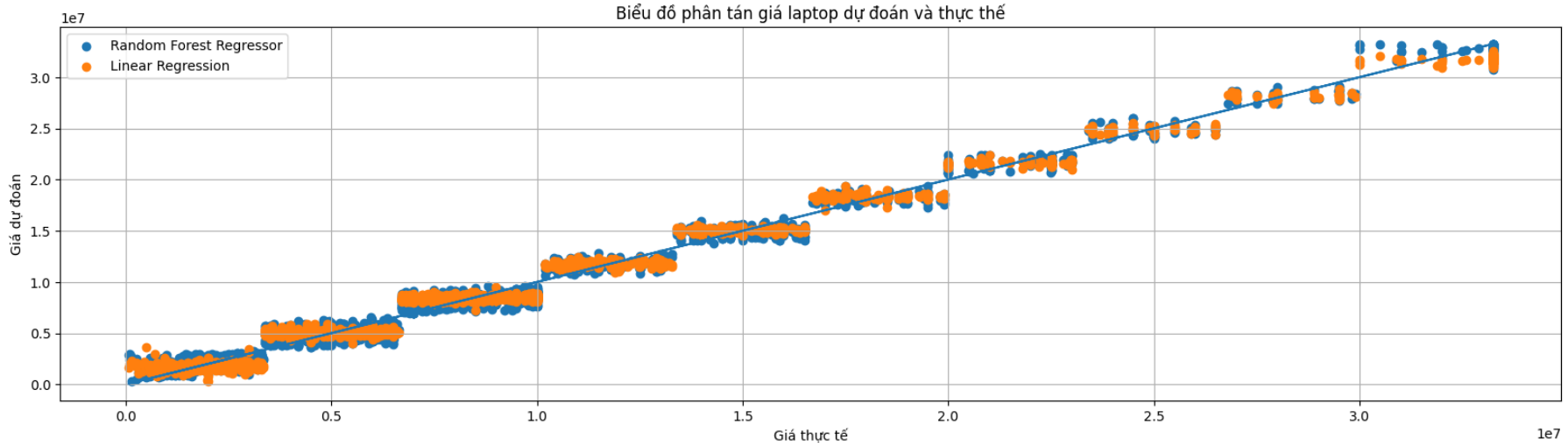




Prediction of laptop's price

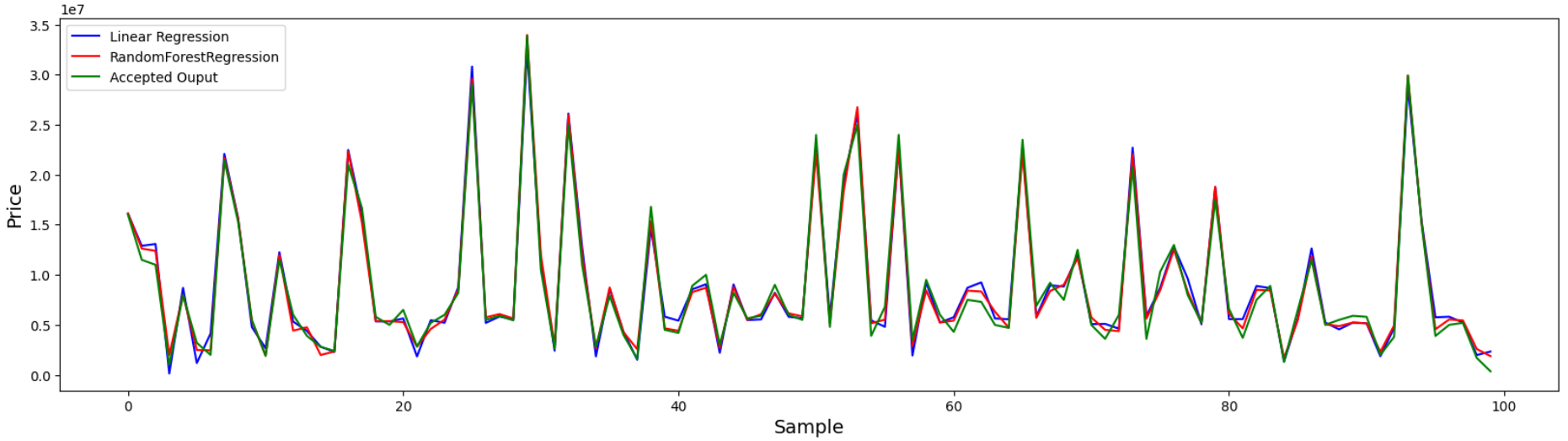
- **Đồ thị kết quả và so sánh 2 mô hình**

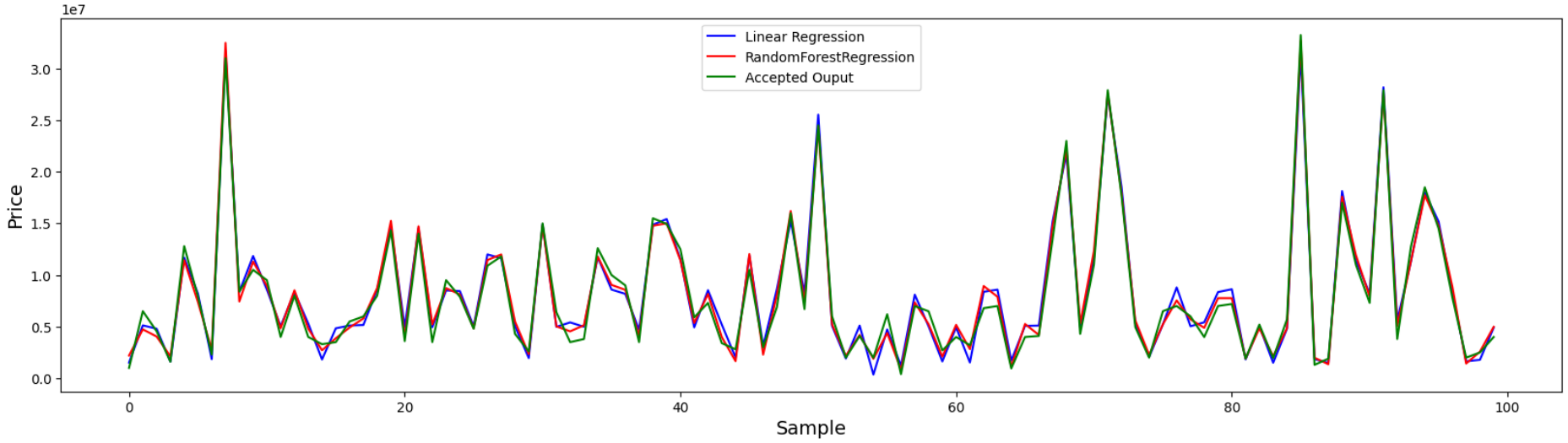




Prediction of laptop's price

- **Đồ thị kết quả và so sánh 2 mô hình**





Prediction of laptop's price

- **Bảng metrics các mô hình**

Small Data

Mô hình	MAE (VND)	RMSE (VND)	MAPE (%)
Random Forest Regressor	753,327.2	929,044.8	14.9
Linear Regression	833,749.8	1015,213.2	16.7

Big Data

Mô hình	MAE (VND)	RMSE (VND)	MAPE (%)
Random Forest Regressor	707,041.0	878,669.2	17.5
Linear Regression	827,058.1	965,991.4	19.6