**Dựđoán giá laptop cũ dựa trên dữ liệu thông số của laptop cũ từ trang thương mại điện tử**

Nguyễn Hùng Trung Hiếu1\*, Trịnh Thế Hiển2\*, Lưu Quang Tiến Hoàng3\*, and Nguyễn Thành Nhân4\*

\*Nhóm 14, *Trường Đại học Công nghệ Thông tin Đại học Quốc Gia – Thành phố Hồ Chí Minh*, Việt Nam

1[20521323@gm.uit.edu.vn](mailto:20521323@gm.uit.edu.vn), 2[20521310@gm.uit.edu.vn](mailto:20521310@gm.uit.edu.vn), 3[20521342@gm.uit.edu.vn](mailto:20521310@gm.uit.edu.vn), 4[20521701@gm.uit.edu.vn](mailto:20521701@gm.uit.edu.vn)

**Tóm tắt.** Hiện nay, thời đại công nghệ ngày càng phát triển mạnh mẽ, dẫn đến mức sống của con người ngày càng cao, các sản phẩm công nghệ tiên tiến phục vụ cho đời sống con người đang nhận được sự ủng hộ và thúc đẩy rất lớn. Đối với học sinh sinh viên ngày nay, chiếc laptop là một thứ không thể thiếu để phục vụ cho học tập và công việc, nhưng không phải ai cũng đủ kinh tế để có thể trang bị những sản phẩm mới, cũng như có đủ chuyên môn để đánh giá một chiếc laptop cũ, do đó trong đề tài này chúng tôi muốn xây dựng bộ dữ liệu phục vụ cho đề tài dự đoán giá laptop cũ. Trong khuôn khổ đồ án này, chúng tôi sử dụng một vài phương pháp xử lý dữ liệu dạng văn bản cũng như vài phương pháp học máy để đánh giá bộ dữ liệu này. Cụ thể, chúng tôi tự xây dựng bộ dữ liệu về thông số cũng như giá cả của những chiếc laptop và xây dựng mô hình học máy đơn giản là linear regression và decision tree để dự đoán với 28 thuộc tính thông số trong bộ dữ liệu.

**Từ khóa:**

# Giới thiệu

## Bối cảnh

Trong thời đại công nghiệp 4.0 hiện nay, máy học trở thành một công nghệ cực kì quan trọng, nó mở đường cho việc tự động hóa nhiều công việc con người không thể làm hoặc mất rất nhiều công sức để làm. Việc tìm hiểu về tựng bộ phận linh kiện và định giá cho một chiếc laptop đã qua sử dụng thì tốn rất nhiều thời gian và công sức, và nó đặc biệt khó đối với nhưng người không có chuyên môn.

Bối cảnh đặt ra là, ngày nay, đối với hầu hết tất cả những học sinh, sinh viên đại học, nhưng người làm công việc văn phòng, quản lý,… thì chiếc laptop là một vật không thể thiếu đối với mỗi người. Từ đó nhu cầu về mua bán, trao đổi laptop là vấn đề ngày càng được mọi người chú ý. Có những người muốn sở hữu những chiếc laptop có cấu hình mạnh nhưng họ lại không đủ kinh tế mua những chiếc máy mới, có những người vì muốn lên đời máy mà muốn bán lại chiếc máy cũ hiện tại của mình để đỡ kinh phí,… nhưng không phải ai cũng có chuyên môn về phần cứng máy tính cũng như định giá một chiếc máy tính cũ đã qua sử dụng. Để có thể thỏa mãn được nhưng thông tin cần thiết, người muốn mua bán, trao đổi có thể phải đến những cửa hàng, trung tâm thiết bị điện tử để nhờ dịch vụ định giá. Và đôi khi vì không có chuyên môn mà những nơi thiếu uy tính có thể lợi dụng và định giá sai cho người dùng.

Từ đó chúng tôi nhận thấy, việc định giá cho những chiếc laptop cũ đem lại lợi ích rất lớn đối với mỗi người trong cuộc sống hiện nay, vì thế chúng tôi quyết định thực hiện đề tài xây dựng bộ dữ liệu dự đoán giá laptop cũ và xây dựng mô hình máy học cho bộ dữ liệu này.

Trong đề tài này, chúng tôi xây dựng bộ dữ liệu về thông số và giá cả của laptop cũ từ trang thương mại điện tử “FPTShop.com.vn” và thực hiện việc đào tạo mô hình máy học đơn giản để đánh giá. Sau khi hoàn thành, chúng tôi có thể phát triển bộ dữ liệu về độ cụ thể và đa dạng, xây dựng và đào tạo mô hình máy học phù hợp với bộ dữ liệu này hơn.

## Định nghĩa bài toán

Task: Dự đoán giá laptop cũ.

E: Gồm 2 tập train và test.

P: Tỉ lệ dự đoán chính xác giá của laptop cũ.

Input: Thông số của một chiếc laptop cũ.

Output: Nhãn giá được dữ đoán cho cho chiếc laptop đó.

Ví dụ:

Input:

Output:

## Mục tiêu

* + Nghiên cứu xây dựng được bộ dữ liệu sạch với độ tin cậy cao
  + Dự đoán được giá laptop cũ với độ chính xác khoảng … %
  + Nghiên cứu và thực hiện được quy trình các bước để xây dựng một bộ dữ liệu cũng như cách xây dựng, huấn luyện và đánh giá mô hình máy học.

## Ứng dụng

## (SUY NGHĨ LÀM CHUNG)

## Công trình liên quan

## (TÌM HIỂU LÀM CHUNG)

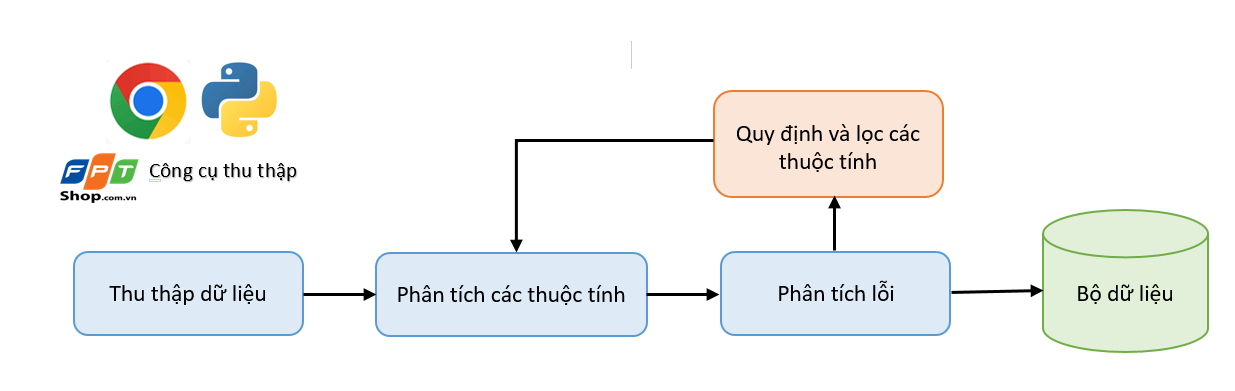
# Dữ liệu

## Thu thập dữ liệu

### 2.1.1 Chuẩn bị

* Chúng tôi tìm hiểu và chuẩn bị một đoạn chương trình để tự động thu thập dữ liệu.
* (HIẾU)

### 2.1.2 Cách thức và quy trình

Chúng tôi đã thu thập dữ liệu về thông số và giá cả của laptop cũ trên trang thương mại điện tử FPT Shop. Sử dụng tiện ích Google Chrome: Google tìm kiếm, Google Colab, Python Visual Code để chạy code thu thập. Chúng tôi thu được khoảng từ 10 – 20 dữ liệu cụ thể về laptop cũ trong 1 lần tải.

*Hình 2.1 Quy trình xây dựng bộ dữ liệu (tên)*

### 2.1.3 Phân tích lỗi

* (HIẾU)

### 2.1.4 Quy định các thuộc tính

* (HIẾU)

## Thông tin về bộ dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Total | Train | Test |
| Sample size | 700 | 560 | 140 |

*Bảng 2.1 Thông tin bộ dữ liệu*

(\*Tên data) là một tập dữ liệu được chúng tôi thu thập trên mạng internet. Cụ thể là thu thập tại nguồn: <https://fptshop.com.vn/>. Tập dữ liệu này chứa 1572 dòng dữ liệu về thông số và giá cả của laptop cũ, với file RawData khi thu thập có đến hơn 70 thuộc tính. Sau khi cân nhắc và để tối ưu hóa khi đưa vào mô hình học máy, File TidyData chỉ giữ lại 703 dòng dữ liệu và 28 thuộc tính.

Link bộ dữ liệu:

<https://drive.google.com/drive/u/4/folders/1hUA49DjfRywdjm4BxGTuzBplYv2Jciw3?fbclid=IwAR0KAQPtMdwlqzh30JbqABsgDIDlplTz4D2_O6LM0cscTEYPE4z8YURsAvA>

## Khảo sát dữ liệu

Tập dữ liệu hoàn chỉnh bao gồm 703 dòng dữ liệu và 28 thuộc tính, thông tin thuộc tính cụ thể như sau:

|  |  |  |
| --- | --- | --- |
| STT | Tên thuộc tính | Kiểu giá trị |
| 1 | Thương hiệu | string |
| 2 | Giá | int |
| 3 | Giá gốc | int |
| 4 | Hạn bảo hành | date |
| 5 | Thiết kế & Trọng lượng.Trọng lượng sản phẩm | float |
| 6 | Thiết kế & Trọng lượng.Chất liệu | string |
| 7 | RAM.Dung lượng RAM | int |
| 8 | RAM.Loại RAM | string |
| 9 | RAM.Số khe RAM còn lại | int |
| 10 | RAM.Hỗ trợ RAM tối đa | int |
| 11 | Bộ xử lý.CPU | string |
| 12 | Bộ xử lý.CPU rank | int |
| 13 | Màn hình.Kích thước màn hình | int |
| 14 | Màn hình.Độ phân giải | string |
| 15 | Màn hình.Tấm nền | string |
| 16 | Màn hình.Độ phủ màu | int |
| 17 | Màn hình.Màn hình cảm ứng | string |
| 18 | Đồ họa.Card onboard | string |
| 19 | Đồ họa.Card rời | string |
| 20 | Lưu trữ.SSD.Dung lượng | int |
| 21 | Lưu trữ.HDD.Dung lượng | int |
| 22 | Lưu trữ.Hỗ trợ công nghệ Optane | string |
| 23 | Giao tiếp & kết nối.Số cổng Thunderbolt | int |
| 24 | Bàn phím & TouchPad.Đèn bàn phím | string |
| 25 | Màn hình.Tần số quét | int |
| 26 | Màn hình.Độ sáng | string |
| 27 | Thông tin pin & Sạc.Dung lượng pin | string |
| 28 | Thông tin pin & Sạc.Tháo rời pin | string |

*Bảng 2.2 Thống kê các thuộc tính*

## Phương pháp xử lý dữ liệu

Đầu tiên, chúng tôi thực hiện lấy dữ liệu tự động của các laptop cũ ở trên FPT Shop thành một file excel. Tiếp đến dử dụng Google Colab để thực hiện thao tác với bộ dữ liệu đã thu thập được.

* Tiến hành upload file excel đã có lên drive.
* Từ Colab sử dụng thư viện thư viện pandas, numpy và re để chuyển file excel đã upload thành file csv để dễ thao tác.
* Dùng pd.read\_csv để tiến hành đọc dữ liệu của file.
* Tiếp theo xóa các cột thuộc tính không cần thiết với bài làm cũng như các cột thuộc tính không quan trọng, những cột thuộc tính mà người dùng ít khi quan tâm khi mua máy (“Thiết kế & Trọng lượng.Kích thước", "Bộ xử lý.Hãng CPU",…).
* Đưa các cột có đơn vị đo lường về dạng số hoặc thuần ký tự (“Nhôm”, “nhựa”,…) để dễ dàng thuận tiện cho việc xử lí dữ liệu sau này.
* Chỉnh sửa các ô dữ liệu NULL trong quá trình lấy dữ liệu về bị thiếu sót.
* Đổi tên các cột bị sai bị thiếu thành các cột mới hoàn chỉnh.
* Tiến hành lọc lại các thuộc tính không cần thiết sau khi đã chỉnh sửa.

Qua đó chúng tôi thu được bộ dữ liệu mới (TinyData) từ bộ dữ liệu cũ (RawData) từ 1572 rows x 78 columns thành 703 rows x 28 columns.

## Đánh giá tập dữ liệu đã xử lý

Dữ liệu có nhiều thuộc tính có kiểu khác nhau, cần được tiền xử lý trước khi thực hiện huấn luyện.

Tập dữ liệu có kích thước nhỏ khi so với các công trình khác, nhưng đủ để thực hiện nghiên cứu trong đồ án này.

Tập train và tập test được chia theo đúng tỉ lệ 8:2.

## Thách thức trong quá trình xây dựng bộ dữ liệu

Trong quá trình xây dựng bộ dữ liệu chúng tôi đã gặp phải một số thách thức như sau:

* Trang web thu thập dữ liệu (fptshop.com.vn) là trang web động (có nhiều button) nên phải sử dụng thêm thư viện.
* Việc thu thập dữ liệu từ nguồn trên internet không đảm bảo sự đồng đều về chất lượng và nhất quán giữa các giá trị có trong thuộc tính khiến quá trình tiền xử lý, làm sạch dữ liệu trở nên khó khăn hơn.
* Có một số lượng dữ liệu bị trùng sau khi tải về từ internet, chúng tôi phải tốn thêm thời gian thực hiện xử lý những dữ liệu trùng lặp.
* Trong quá trình tải, có một số lượng dữ liệu bị lẫn dữ liệu rác, link quảng cáo là những dữ liệu không thuộc trong đề tài này.

# Phương pháp máy học

## Mô hình máy học

### 3.1.1 Mô hình …

### 3.1.2 Mô hình …

### 3.1.3 Mô hình …

## Công cụ sử dụng

## (Mình thêm bớt gì sửa đoạn mẫu bên dưới nha)

Nền tảng sử dụng: Google Colab.

Thư viện sử dụng:

Sklearn: là thư viện cung cấp một tập các công cụ xử lý các bài toán machine learning và statistical modeling.

Pandas: là thư viện dùng để thao tác, phân tích và dọn dẹp dữ liệu.

Numpy: là thư viện chúng tôi dùng để xử lý mảng đa chiều, ma trận.

Cv2: là thư viện cho phép đọc, ghi, thay đổi dữ liệu nhiều hình ảnh cùng một lúc.

Math: module này cho phép truy cập vào các hàm lượng giác, hàm số, hàm logarit cho các số thực.

Os: module này cho phép thao tác với tệp và thư mục.

Random: module cho phép tạo ra một số ngẫu nhiên bất kì với nhiều yêu cầu khác nhau.

Matplotlib: là thư viện dùng để vẽ đồ thị 2D.

Seaborn: là một thư viện trực quan hóa dữ liệu.

PIL: có thể mở, lưu, xử lý đặc điểm hình ảnh với nhiều định dạng ảnh khác nhau.

Skimage: là một thư viện xử lý ảnh nguồn mở bao gồm các thuật toán để phân đoạn, thao tác không gian màu, phân tích, lọc, phát hiện tính năng, ...

Công cụ khác:

GridSearchCV: Lấy một từ điển mô tả các tham số có thể được thử trên một mô hình để huấn luyện nó. Lưới tham số được định nghĩa như một từ điển, trong đó các khóa là các tham số và các giá trị là cài đặt cần kiểm tra.

## Các phương pháp đánh giá

## (Mình thêm bớt gì sửa đoạn mẫu bên dưới nha)

Confusion matrix: Giúp đánh giá được các giá trị cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất, và dữ liệu thuộc lớp nào hay bị phân loại nhầm vào lớp khác.

True Positive (TP): Giá trị dự đoán khớp với giá trị thực tế. Nhãn dương tính được dự đoán là dương tính.

True Negative (TN): Giá trị dự đoán khớp với giá trị thực tế. Nhãn âm tính được dự đoán là âm tính.

False Positive (FP): Giá trị dự đoán bị sai. Nhãn âm tính bị dự đoán là dương tính.

False Negative (FN): Giá trị dự đoán bị sai. Nhãn dương tính bị dự đoán là âm tính.

Chúng tôi sử dụng 2 độ đo sau đây để đánh giá mô hình:

### 3.3.1 F1-score:

F1-score là trung bình điều hòa của recall và precision. F1-score nằm trong khoảng (0,1] F1 càng cao, bộ phân loại càng tốt.

F1=2\*(Precision\*Recall)/(Precision+Recall)

Trong đó:

Recall=TP/(TP+FN)

Precision=TP/(TP+FP)

Bản chất bài toán phân loại loài vật là phân lớp đa lớp với các nhãn có trọng số như nhau, do vậy recall cao mà precision thấp thì không phải mô hình tốt và ngược lại, precision cao mà recall thấp cũng không phải mô hình tốt. Do vậy chúng tôi quyết định sử dụng F1-score.

### 3.3.2 Accuracy

Accuracy đánh giá mô hình bằng cách tính tỉ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử.

Accuracy=(TP +TN)/(TP+TN+FP+FN)

Đề tài của chúng tôi có điểm dữ liệu của các nhãn bằng nhau, vậy nên chúng tôi dùng thêm Accuracy vì đây là độ đo rất phù hợp với bài toán phân lớp có bộ dữ liệu cân bằng.

## Các thử nghiệm tinh chỉnh mô hình

* 1. **Hướng phát triển**

# Kết luận

## Kết quả đạt được

Đồ án đạt mục tiêu đã được đề ra từ đầu, đó là xây dựng được một mô hình cho kết quả đạt độ chính xác khoảng 70% (SVC đạt 70%)

Tìm hiểu, sử dụng được một số phương pháp tiền xử lý ảnh có thể ứng dụng trong các bài toán tương tự.

Tìm hiểu và tiếp cận được nhiều phương pháp máy học cho bài toán phân lớp đa lớp. Có cái nhìn cơ bản tổng thể về Machine Learning.

## Khó khăn gặp phải

Thiếu kinh nghiệm giải quyết các bài toán xử lý dữ liệu có nhiều thuộc tính nói riêng và các bài toán Machine Learning nói chung dẫn đến việc gặp nhiều khó khăn trong quá trình thu thập cũng như xử lý dữ liệu.

Đã sử dụng nhiều phương pháp xử lý dữ liệu nhưng một số lỗi vẫn không cho kết quả khả quan.

Hạn chế về kiến thức nền khiến cho việc tiếp cận các phương pháp xử lý dữ liệu cũng như tinh chỉnh mô hình trong python tương đối khó khăn.

# Tài liệu tham khảo

1. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197. doi:10.1016/0022-2836(81)90087-5
2. May P, Ehrlich H-C, Steinke T (2006) ZIB structure prediction pipeline: composing a complex biological workflow through web services. In: Nagel WE, Walter WV, Lehner W (eds) Euro-Par 2006. LNCS, vol 4128. Springer, Heidelberg, pp 1148–1158. doi:10.1007/11823285\_121
3. Foster I, Kesselman C (1999) The grid: blueprint for a new computing infrastructure. Mor- gan Kaufmann, San Francisco
4. Czajkowski K, Fitzgerald S, Foster I, Kesselman C (2001) Grid information services for distributed resource sharing. In: 10th IEEE international symposium on high performance distributed computing. IEEE Press, New York, pp 181–184.

doi:10.1109/HPDC.2001.945188

1. Foster I, Kesselman C, Nick J, Tuecke S (2002) The physiology of the grid: an open grid services architecture for distributed systems integration. Technical report, Global Grid Fo- rum
2. National Center for Biotechnology Information. [http://www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov/)