TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

**KHOA ĐIỆN TỬ**

**Bộ môn: Công nghệ thông tin**

****

**BÀI TẬP LỚN**

MÔN HỌC

**KHOA HỌC DỮ LIỆU**

**Sinh viên : Hoàng Thị Hường**

**Lớp : K55KMT**

**Giáo viên hướng dẫn : Nguyễn Văn Huy**

**Thái Nguyên – 2023**

|  |  |  |  |
| --- | --- | --- | --- |
| **TRƯỜNG ĐHKTCN** | | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM** | |
| **KHOA ĐIỆN TỬ** | | ***Độc lập - Tự do - Hạnh phúc*** | |

**BÀI TẬP LỚN**

**MÔN HỌC: KHOA HỌC DỮ LIỆU**

BỘ MÔN : CÔNG NGHỆ THÔNG TIN

*Sinh viên:Hoàng Thị Hường. MSSV: K195480106013*

*Lớp*: K55KMT *Ngành: Kỹ Thuật Máy Tính*

*Ngày giao đề: ../05/2023. Ngày hoàn thành: 30/05/2023*

*Tên đề tài: Dự đoán giá điện thoại*

*Yêu cầu*

|  |
| --- |
| **GIÁO VIÊN HƯỚNG DẪN** |
| *(Ký và ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

Thái Nguyên, ngày….tháng…..năm 20....

GIÁO VIÊN HƯỚNG DẪN

*(Ký ghi rõ họ tên)*

**NHẬN XÉT CỦA GIÁO VIÊN CHẤM**

**………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………**

*Thái Nguyên, ngày … tháng … năm ……*

**GIÁO VIÊN CHẤM**

*(Ký và ghi rõ họ tên)*

MỤC LỤC

[LỜI NÓI ĐẦU 6](#_Toc136417565)

[CHƯƠNG I. GIỚI THIỆU 7](#_Toc136417566)

[1.1. Đoán giá là gì? 7](#_Toc136417567)

[1.2. Ứng dụng 8](#_Toc136417568)

[1.3. Công nghệ 9](#_Toc136417569)

[1.4. Áp dụng 10](#_Toc136417570)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 11](#_Toc136417571)

[2.1. Dữ liệu 11](#_Toc136417572)

[*2.1.1. Thu thập dữ liệu* 11](#_Toc136417573)

[*2.1.2. Xử lý dữ liệu* 12](#_Toc136417574)

[2.2. Model 15](#_Toc136417575)

[CHƯƠNG 3. XÂY DỰNG CHƯƠNG TRÌNH 17](#_Toc136417576)

[3.1. Chương trình dự đoán 17](#_Toc136417577)

[CHƯƠNG 4. KẾT LUẬN 19](#_Toc136417578)

[1. Tổng quan 19](#_Toc136417579)

[2. Ưu, nhược điểm 19](#_Toc136417580)

[a) Ưu điểm 19](#_Toc136417581)

[b) Nhược điểm. 19](#_Toc136417582)

[TÀI LIỆU THAM KHẢO 20](#_Toc136417583)

# LỜI NÓI ĐẦU

Điện thoại di động đã trở thành một phần không thể thiếu trong cuộc sống hiện đại của chúng ta. Với sự phát triển nhanh chóng của công nghệ, thị trường điện thoại di động cũng ngày càng đa dạng và cạnh tranh. Một trong những yếu tố quan trọng khi người dùng đưa ra quyết định mua điện thoại mới là giá cả.

Dự đoán giá điện thoại là một vấn đề quan trọng và hấp dẫn trong lĩnh vực kinh tế và công nghệ. Việc hiểu và dự đoán giá điện thoại có thể mang lại lợi ích lớn cho các nhà sản xuất, nhà bán lẻ và người tiêu dùng. Đối với các nhà sản xuất, việc định giá sản phẩm một cách chính xác và hợp lý có thể giúp họ nắm bắt được xu hướng thị trường và phát triển chiến lược kinh doanh hiệu quả. Đối với nhà bán lẻ, việc đoán giá đúng có thể tăng tính cạnh tranh và thu hút khách hàng. Còn đối với người tiêu dùng, dự đoán giá điện thoại có thể giúp họ đưa ra quyết định mua hàng thông minh và tiết kiệm.

# CHƯƠNG I. GIỚI THIỆU

## Đoán giá là gì?



Hình 1.1. Đoán giá là gì?

Đoán giá (price prediction) là quá trình dự đoán giá trị của một tài sản, sản phẩm hoặc dịch vụ dựa trên các thông tin và dữ liệu hiện có. Mục đích của việc đoán giá là cung cấp cho người dùng một sự ước lượng về giá trị tiềm năng trong tương lai, giúp họ đưa ra quyết định đầu tư hoặc mua bán dựa trên thông tin đó.

Đoán giá có thể được áp dụng trong nhiều lĩnh vực khác nhau, bao gồm thị trường chứng khoán, thị trường tiền điện tử, thị trường bất động sản, ngành công nghiệp, và nhiều lĩnh vực kinh tế khác. Có nhiều phương pháp và kỹ thuật khác nhau được sử dụng để đoán giá, bao gồm phân tích kỹ thuật, phân tích cơ bản, mô hình hóa thống kê, trí tuệ nhân tạo và học máy. Các yếu tố quan trọng được sử dụng để đưa ra dự đoán bao gồm các dữ liệu lịch sử, chỉ số kỹ thuật, tin tức và yếu tố tác động đến thị trường hoặc ngành công nghiệp liên quan.

Tuy nhiên, đoán giá không phải là một công việc chính xác 100% và có thể gặp phải sai sót. Nhiều yếu tố có thể ảnh hưởng đến giá trị thực tế, bao gồm biến động thị trường, tác động của sự kiện không lường trước, thông tin mới và các yếu tố khác mà các mô hình dự đoán không thể hiểu được. Do đó, đoán giá thường mang tính chất tham khảo và không đảm bảo đúng đắn.

## 1.2. Ứng dụng

Có nhiều ứng dụng của việc đoán giá trong các lĩnh vực khác nhau. Dưới đây là một số ví dụ phổ biến về các ứng dụng của đoán giá:

* Tài chính và chứng khoán: Trong lĩnh vực này, dự đoán giá được sử dụng để đưa ra quyết định đầu tư thông minh và tối ưu. Các nhà đầu tư và các công ty tài chính sử dụng dự đoán giá để đánh giá rủi ro và tiềm năng sinh lợi từ các tài sản như cổ phiếu, quỹ đầu tư và các loại chứng khoán khác.
* Bất động sản: Dự đoán giá được áp dụng để ước tính giá trị của bất động sản như nhà ở, căn hộ, đất đai và tài sản thương mại. Những thông tin này có thể hỗ trợ các nhà đầu tư, môi giới và các bên liên quan khác trong việc đưa ra quyết định mua bán, định giá và quản lý tài sản bất động sản.
* Thương mại điện tử: Trong lĩnh vực thương mại điện tử, việc đoán giá có thể được sử dụng để ước lượng giá sản phẩm và dịch vụ trên các nền tảng mua sắm trực tuyến. Các công ty thương mại điện tử thường sử dụng các thuật toán dự đoán giá để đề xuất giá cả cạnh tranh và tùy chỉnh cho khách hàng.
* Du lịch và khách sạn: Trong ngành du lịch và khách sạn, việc đoán giá có thể được sử dụng để ước lượng giá phòng khách sạn, vé máy bay, tour du lịch và các dịch vụ du lịch khác. Các công ty du lịch, đại lý du lịch và các trang web đặt phòng sử dụng các mô hình đoán giá để quản lý giá cả, tối ưu hóa doanh thu và cung cấp ưu đãi cho khách hàng.

## 1.3. Công nghệ

Học máy (Machine Learning): Học máy là một phương pháp sử dụng các thuật toán và mô hình để học từ dữ liệu và dự đoán kết quả trong tương lai. Các thuật toán phổ biến trong lĩnh vực đoán giá bao gồm hồi quy tuyến tính, hồi quy logistic, máy vector hỗ trợ (SVM), cây quyết định và các thuật toán học sâu như mạng neural.

Học sâu (Deep Learning): Học sâu là một lĩnh vực của trí tuệ nhân tạo (AI) tập trung vào việc xây dựng và huấn luyện các mạng neural sâu để hiểu và giải quyết các bài toán phức tạp. Trong lĩnh vực đoán giá, mạng neural sâu, chẳng hạn như mạng neural tích chập (CNN) hoặc mạng neural hồi quy (RNN), có thể được sử dụng để học từ dữ liệu và dự đoán giá trị.

Kỹ thuật tăng cường (Ensemble Techniques): Kỹ thuật tăng cường là một phương pháp kết hợp nhiều mô hình dự đoán để tạo ra một dự đoán tốt hơn. Ví dụ: Kỹ thuật Bagging (Bootstrap Aggregating) sử dụng nhiều mô hình học máy độc lập và kết hợp kết quả để đưa ra dự đoán cuối cùng.

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP): Trong một số trường hợp, đoán giá có thể dựa trên thông tin và mô tả liên quan đến sản phẩm hoặc dịch vụ. Xử lý ngôn ngữ tự nhiên là một lĩnh vực của AI tập trung vào việc xử lý và hiểu ngôn ngữ tự nhiên của con người. Trong đoán giá, NLP có thể được sử dụng để phân tích và rút trích thông tin từ các mô tả sản phẩm hoặc nhận xét của người dùng để đưa ra dự đoán về giá trị.

Kỹ thuật trích xuất đặc trưng (Feature Extraction): Trích xuất đặc trưng là quá trình chuyển đổi dữ liệu đầu vào thành một tập hợp các đặc trưng có ý nghĩa và dễ dùng cho việc dự đoán. Trong đoán giá, kỹ thuật trích xuất đặc trưng có thể bao gồm việc chuyển đổi dữ liệu định tính thành dữ liệu số, trích xuất thông tin từ hình ảnh hoặc văn bản, và xử lý các biến đầu vào để tạo ra các đặc trưng mới..

Những công nghệ này có thể được kết hợp và tùy chỉnh phù hợp với bài toán đoán giá cụ thể để đạt được kết quả tốt nhất. Việc lựa chọn công nghệ phụ thuộc vào tính chất của dữ liệu, độ phức tạp của bài toán, và tài nguyên có sẵn.

## 1.4. Áp dụng

Trong bài em có sử dụng dữ liệu từ kaggle và dùng mô hình hồi quy tuyến tính(Linear regression) để dự đoán giá điện thoại dựa trên các đặc trưng như kích thước, thương hiệu, hệ điều hành, dung lượng bộ nhớ, RAM, năm sản xuất, v.v.

Hồi quy tuyến tính được sử dụng để tìm hiểu mối quan hệ tuyến tính giữa các đặc trưng của điện thoại và giá, và từ đó, mô hình có thể dự đoán giá của các mẫu mới dựa trên các đặc trưng đó.

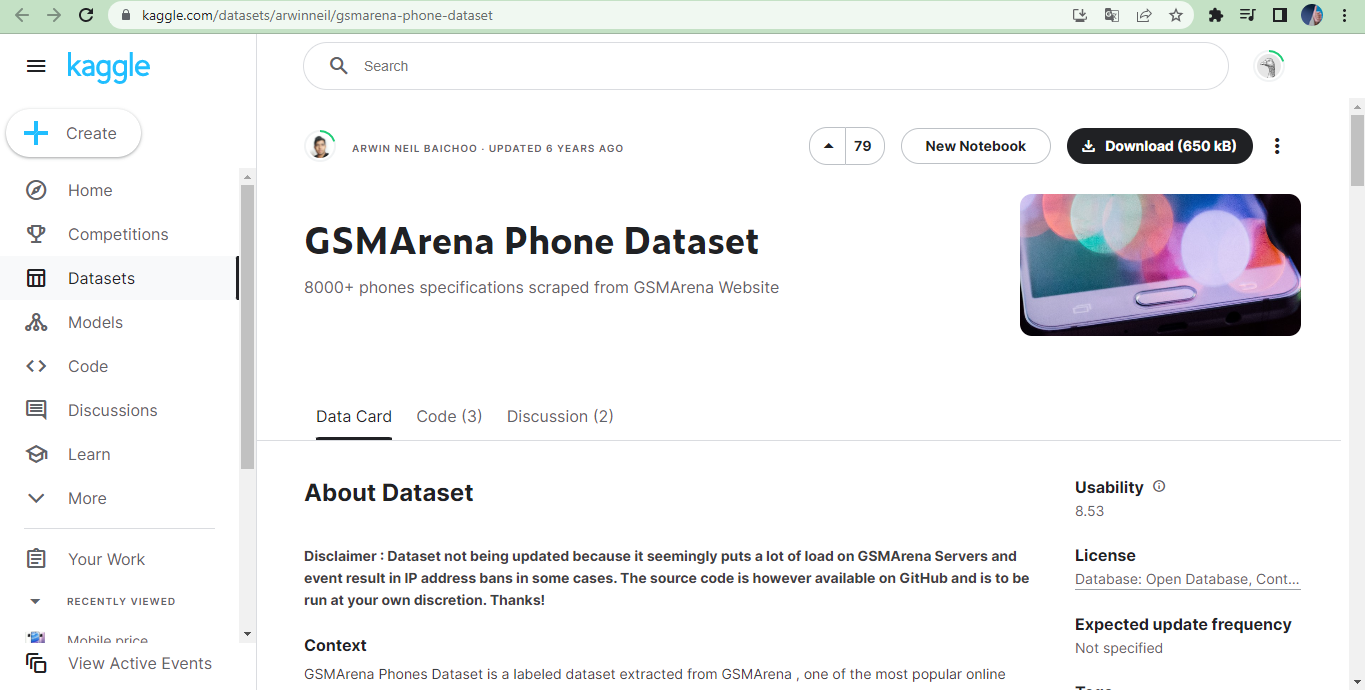
# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

## 2.1. Dữ liệu

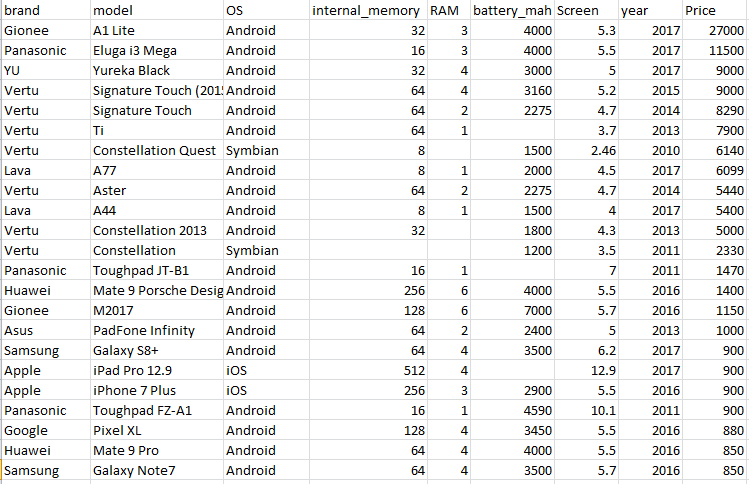
### *2.1.1. Thu thập dữ liệu*

Trước khi xử lý dữ liệu, chúng ta cần thu thập dữ liệu về các mẫu điện thoại đã được ra mắt trước đó. Báo cáo này sử dụng bộ dữ liệu "GSMArena Phone Dataset" từ trang web Kaggle. Bộ dữ liệu này cung cấp thông tin về các điện thoại di động, bao gồm thông số kỹ thuật và giá bán.

Bộ dữ liệu "GSMArena Phone Dataset" chứa thông tin về hàng ngàn mẫu điện thoại từ nhiều nhà sản xuất khác nhau. Các thông tin chi tiết bao gồm thông số kỹ thuật như kích thước màn hình, dung lượng pin, RAM, bộ nhớ trong, hệ điều hành và nhãn hiệu. Ngoài ra, bộ dữ liệu cũng cung cấp thông tin về giá bán của các mẫu điện thoại này.



Hình 2.1. Bộ dữ dữ liệu trên trang Kaggle.



Hình 2.2. Dữ liệu data\_phone.csv sau khi tải về.

Dữ liệu này bao gồm các thuộc tính:

- brand : Hãng điện thoại

- model : Tên điện thoại

- OS : Hệ điều hành

- internal\_memory : bộ nhớ trong(GB)

- RAM : dung lượng ram(GB)

- battery : dung lượng pin(mAh)

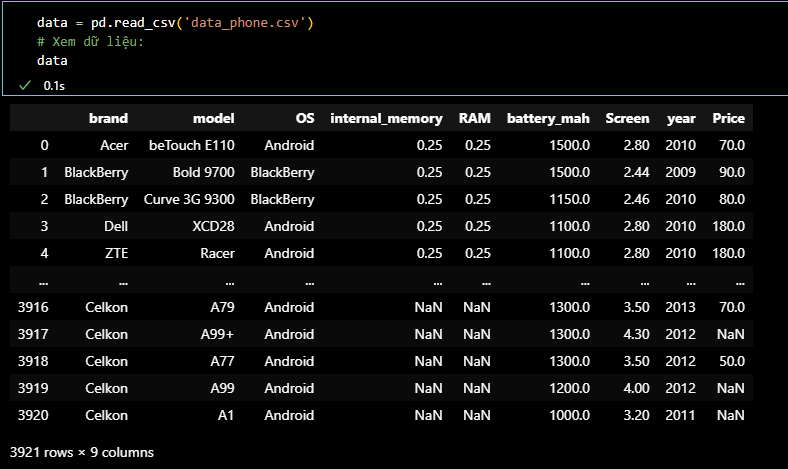
- Screen : kích thước màn hình(inches)

- year  : năm sản xuất

- Price : giá của điện thoại(EUR)

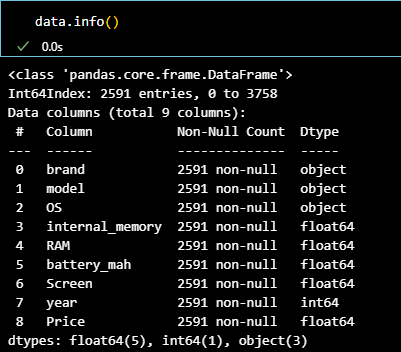
### *2.1.2. Xử lý dữ liệu*

Bước tiếp theo sau khi thu thập được bộ dữ liệu về các mẫu điện thoại là xử lý dữ liệu để chuẩn bị cho quá trình dự đoán giá điện thoại. Quá trình này bao gồm loại bỏ các dữ liệu rỗng, xử lý nhiễu và chuẩn hóa dữ liệu.



Hình 2.3. Đọc file csv bằng thư viện pandas.

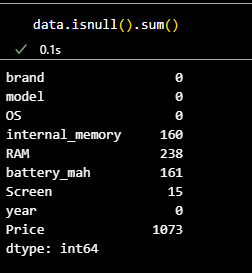
Quá trình bắt đầu bằng việc đọc file csv chứa dữ liệu. Sau khi xem qua dữ liệu đầu vào, xem xét các thuộc tính và thông số khác nhau để đánh giá dữ liệu và xác định những bước cần thực hiện để chuẩn bị dữ liệu cho mô hình.



Hình 2.4. Thông tin kiểu dữ liệu.

Dữ liệu đầu vào bao gồm ba loại kiểu dữ liệu: float64, int64 và object. Kiểu dữ liệu float là kiểu dữ liệu số dạng thập phân, int là kiểu dữ liệu số dạng số nguyên, và object đại diện cho kiểu dữ liệu gốc hoặc không rõ ràng.

Trong bộ dữ liệu, có ba trường thuộc kiểu object là "brand", "model" và "os". Những trường này chứa thông tin về thương hiệu, mẫu điện thoại và hệ điều hành. Kiểu dữ liệu object thường được hiểu là kiểu dữ liệu văn bản hoặc chuỗi ký tự, trong đó giá trị có thể là tên thương hiệu, tên mẫu điện thoại và tên hệ điều hành.

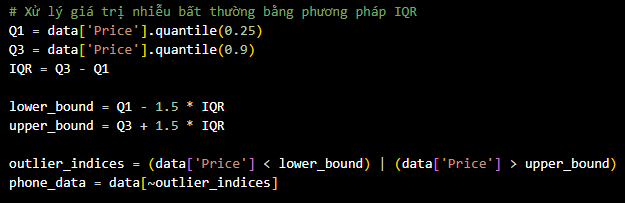


Hình 2.5. Kiểm tra dữ liệu rỗng(NaN).

Dữ liệu rỗng có thể gây nhiễu và ảnh hưởng đến tính toàn vẹn và chất lượng của mô hình dự đoán. Tất cả các giá trị rỗng trong bộ dữ liệu cần được xử lý để đảm bảo tính toàn vẹn và chất lượng của dữ liệu.

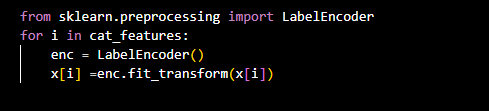
Phương pháp phổ biến để xử lý dữ liệu rỗng là loại bỏ các mẫu dữ liệu chứa giá trị rỗng. Điều này có thể được thực hiện bằng cách sử dụng phương thức dropna() trên bộ dữ liệu. Kết quả trả về sẽ là một bộ dữ liệu mới chỉ chứa các mẫu dữ liệu hoàn chỉnh.

Sau khi hoàn thành quá trình xử lý dữ liệu bị thiếu, tiếp theo chúng ta sẽ tiến hành xử lý các giá trị nhiễu trong dữ liệu đầu vào. phương pháp phổ biến để xử lý giá trị nhiễu là sử dụng phương pháp IQR (Interquartile Range). Phương pháp này giúp chúng ta xác định và loại bỏ các giá trị nhiễu dựa trên sự khác biệt giữa phần vị vành đai thứ nhất (Q1) và phần vị vành đai thứ ba (Q3) của phân phối dữ liệu.



Hình 2.6. Xử lý nhiễu bằng IQR.

Tiếp theo chúng ta cần xử lý dữ liệu kiểu object. Trong trường hợp này, một phương pháp phổ biến là sử dụng hàm LabelEncoder() để mã hoá các biến thành các số nguyên. Mục đích của việc này là chuyển đổi dữ liệu văn bản thành dạng số để sử dụng cho việc xây dựng mô hình dự đoán.

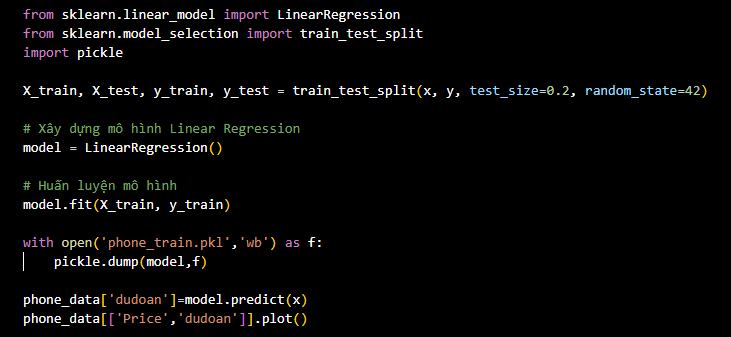


Hình 2.7. Mã hóa dữ liệu object.

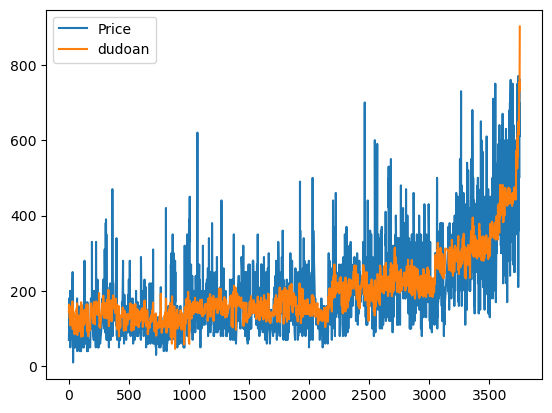
## 2.2. Model

Trong đề tài này, mô hình Linear Regression được sử dụng để dự đoán giá điện thoại.

Linear regression (Hồi quy tuyến tính) là một thuật toán trong lĩnh vực học máy được sử dụng để xây dựng mô hình dự đoán dựa trên mối quan hệ tuyến tính giữa biến đầu vào (biến độc lập) và biến đầu ra (biến phụ thuộc). Mô hình hồi quy tuyến tính giả định rằng có một mối quan hệ tuyến tính giữa các biến đầu vào và đầu ra.



Hình 2.8 xây dựng mô hình bằng phương pháp Linear Regression

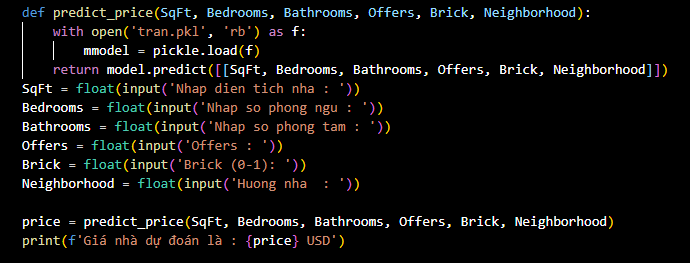


Hình 2.9. Sơ đồ dự đoán so với dữ liệu ban đầu.

Nhìn qua biểu đồ có thể thấy được dự đoán có kết quả không được khả quan lắm.

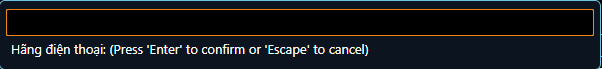
# CHƯƠNG 3. XÂY DỰNG CHƯƠNG TRÌNH

## 3.1. Chương trình dự đoán

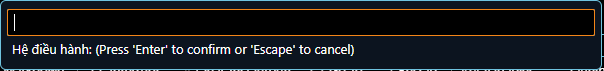


Hình 3.1. chương trình dự đoán.

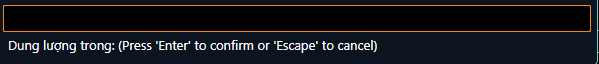
3.2. Chạy chương trình



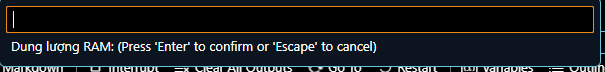
Sau khi chạy chương trình, màn hình sẽ hiện lên khung text để người dùng có thể nhập dữ liệu từ bàn phím. Đầu tiên, nhập hãng điện thoại.

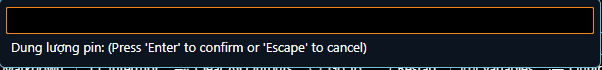


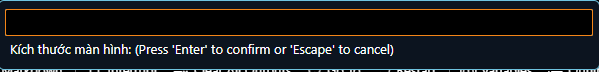
Sau khi ấn Enter, tiếp tục nhập hệ điều hành.

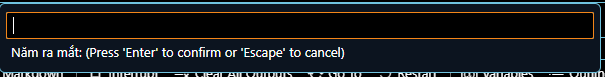


Sau đó nhập dung lượng bộ nhớ trong.



Tiếp đến, ta nhập dung lượng ram. 

Nhập dung lượng pin.

Nhập kích thước màn hình.

Nhập năm sản xuất .



Hệ thống sẽ chạy và đưa ra kết quả dự đoán giá.

# CHƯƠNG 4. KẾT LUẬN

1. Tổng quan

Chương trình đã có thể thu thập các thông tin được nhập qua giao diện rồi đưa vào model đã được xây dựng để dự đoán giá điện thoại.

2. Ưu, nhược điểm

1. Ưu điểm

* Giao diện thân thiện dễ sử dụng.
* Tốc độ dự đoán nhanh chóng, không bị gián đoạn.
* Dữ liệu đã được xử lý và chuẩn hóa một cách hiệu quả, giúp cải thiện khả năng dự đoán.

1. Nhược điểm.

* Độ sai lệch vẫn còn lớn.
* Dữ liệu đầu vào có thể không đầy đủ hoặc chưa hoàn thiện, có thể gây ảnh hưởng đến độ chính xác của mô hình dự đoán..

# TÀI LIỆU THAM KHẢO

[1]. <https://www.kaggle.com/code/hypopossum/gsm-arena-phone-dataset/input>