**TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP THÁI NGUYÊN**

**KHOA: ĐIỆN TỬ**

**BỘ MÔN: TIN HỌC CÔNG NGHIỆP**

****

**BÀI TẬP LỚN**

**MÔN HỌC**

**KHOA HỌC DỮ LIỆU**

**Sinh viên: Lê Thu Phương**

**Lớp: K55KMT.01**

**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: Lê Thu Phương MSSV: K195480106017*

*Lớp: K55KMT.01*

*Ngành: Kỹ Thuật Máy Tính*

*Ngày giao đề:25/04/2023 Ngày hoàn thành:25/05/2023*

1. Tên đề tài: Dự đoán giá xe máy đã sử dụng.

2. Yều cầu:

Thu thập, phân tích dữ liệu

Sử dụng mô hình hồi quy tuyến tính để train, test mô hình

Xây dựng app và dự đoán giá xe máy.

TRƯỞNG BỘ MÔN GIÁO VIÊN HƯỚNG DẪN

*(Ký và ghi rõ họ tên) (Ký và ghi rõ họ tên)*

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

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

*Thái Nguyên, ngày … tháng … năm ……*

**GIÁO VIÊN HƯỚNG DẪN**

*(Ký và 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 HÌNH ẢNH**

Hình 1.Đoán giá

Hình 2: Dữ liệu thu thập được

Hình 3 : Data trên python

Hình 4: Thông tin dữ liệu

Hình 5 :Mô hình

Hình 6 :Kết quả R2

Hình 7 :Chương trình

Hình 8 :Giao diện chương trình

# **MỤC LỤC**

[**MỤC LỤC** 6](#_Toc136426406)

[**LỜI NÓI ĐẦU** 8](#_Toc136426407)

[**CHƯƠNG 1. GIỚI THIỆU CHUNG** 9](#_Toc136426408)

[**1.Khái niệm đoán giá** 9](#_Toc136426409)

[**2.Ứng dụng** 10](#_Toc136426410)

[**3.Công nghệ** 11](#_Toc136426411)

[**4. Áp dụng trong đê tài** 12](#_Toc136426412)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT** 13](#_Toc136426413)

[**2.1. Xử lý data** 13](#_Toc136426414)

[*2.1.1.Thu thập dữ liệu* 13](#_Toc136426415)

[*2.1.2.Các data cần xử lý(làm sạch)* 13](#_Toc136426416)

[*2.1.3.Xử lý data với python* 15](#_Toc136426417)

[**2.2. Model** 16](#_Toc136426418)

[**CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH** 17](#_Toc136426419)

[**3.1.Thuật toán** 17](#_Toc136426420)

[**3.2.Xây dựng chương trình** 18](#_Toc136426421)

[**CHƯƠNG 4: KẾT LUẬN** 19](#_Toc136426424)

[**TÀI LIỆU THAM KHẢO** 21](#_Toc136426425)

# **LỜI NÓI ĐẦU**

Ngày nay việc sở hữu một chiếc xe máy để làm phương tiện đi lại đã trở nên phổ biến hơn bao giờ hết. Tuy nhiên, khi muốn bán lại chiếc xe đã qua sử dụng, chúng ta thường gặp khó khăn trong việc định giá xe một cách chính xác và hợp lý. Thậm chí, chúng ta có thể bị ép giá, khi không có đủ kiến thức về xe. Vì vậy, việc xây dựng một hệ thống dự đoán giá xe máy trở thành một nhu cầu thiết yếu để giúp chúng ta đánh giá giá trị của chiếc xe một cách chính xác và công bằng.

Để xây dựng một hệ thống dự đoán giá xe cũ đáp ứng được yêu cầu của người dùng là một thách thức. Để dự đoán giá trị của một chiếc xe cũ, chúng ta cần phải xem xét nhiều yếu tố khác nhau, như tuổi của xe, số dặm đã đi, tình trạng và chất lượng xe, vị trí địa lý, giá bán trên thị trường...và nhiều yếu tố khác. Do đó, việc tạo ra một hệ thống dự đoán giá xe cũ chính xác đòi hỏi sự phân tích kỹ lưỡng.

Từ nhu cầu thực tế đó em quyết định chọn đề tài xây dựng hệ thống dự đoán giá xe máy đã sử dụng làm đề tài bài tập lớn của mình. Người dùng cần có thể truy cập hệ thống dễ dàng và nhanh chóng và nhận được kết quả dự đoán giá xe một cách chính xác và tin cậy.

# **CHƯƠNG 1. GIỚI THIỆU CHUNG**

## **1.Khái niệm đoán giá**



*Hình 1.Đoán giá*

Đoán giá (price prediction) là quá trình ước tính giá trị tương lai của một tài sản, sản phẩm hoặc dịch vụ dựa trên các thông tin hiện có và các phương pháp phân tích dữ liệu. Nó thường được áp dụng trong lĩnh vực tài chính, thị trường chứng khoán, thương mại điện tử và nhiều lĩnh vực khác để đưa ra các dự đoán về giá cả.

Để đoán giá , các chuyên gia thường sử dụng các kỹ thuật và công cụ trong lĩnh vực khoa học dữ liệu và học máy như mô hình hồi quy, mạng neural, phân tích chuỗi thời gian và các thuật toán khác. 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, để có dự đoán giá chính xác, thông tin cụ thể về sản phẩm hoặc dịch vụ cần được xem xét, cũng như các yếu tố kinh tế và thị trường liên quan. Trong một số trường hợp, các mô hình dự đoán hoặc thuật toán máy học có thể được sử dụng để phân tích dữ liệu và đưa ra dự đoán về giá trị trong tương lai.

## **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:

* Bất động sản: Đoán giá cũng được áp dụng trong lĩnh vực bất động sản để ước lượng giá trị các tài sản như căn hộ, nhà đất hoặc tòa nhà. Các nhà phát triển, nhà môi giới và các chuyên gia bất động sản sử dụng các mô hình đoán giá để đưa ra quyết định về mua, bán hoặc đầu tư vào 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.
* Ngành y tế: Đoán giá cũng có thể được áp dụng trong ngành y tế để ước lượng giá trị của các dịch vụ y tế và sản phẩm dược phẩm. Các công ty bảo hiểm y tế, bệnh viện và nhà nghiên cứu y tế sử dụng các mô hình đoán giá để đưa ra quyết định về giá cả, chi phí và quản lý tài chính.
* Tài chính và chứng khoán: Trong lĩnh vực tài chính, việc đoán giá có thể được sử dụng để dự đoán giá cổ phiếu, hàng hóa, ngoại tệ và các tài sản tài chính khác. Các nhà đầu tư và các công ty tài chính thường sử dụng các mô hình và thuật toán dự đoán để đưa ra quyết định giao dịch và đầu tư.

## **3.Công nghệ**

Công nghệ đóng vai trò quan trọng trong việc đoán giá và các ứng dụng liên quan. Dưới đây là một số công nghệ thường được sử dụng trong việc đoán giá:

* 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..

Các 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.

## **4. Áp dụng trong đê tài**

Trong đề tài em làm, em có sử dụng dữ liệu từ [kaggle](https://www.kaggle.com/code/venkatganesh98/old-car-price-prediction-randomforestregr-acc-85/notebook#3.Data-Pre-Processing). Và dùng mô hình hồi quy tuyến tính (Linear Regression) được sử dụng để dự đoán giá của xe dựa trên các đặc trưng như tên xe, công ty sản xuất, năm sản xuất, số kilomet đã đi và động cơ, công suất,...

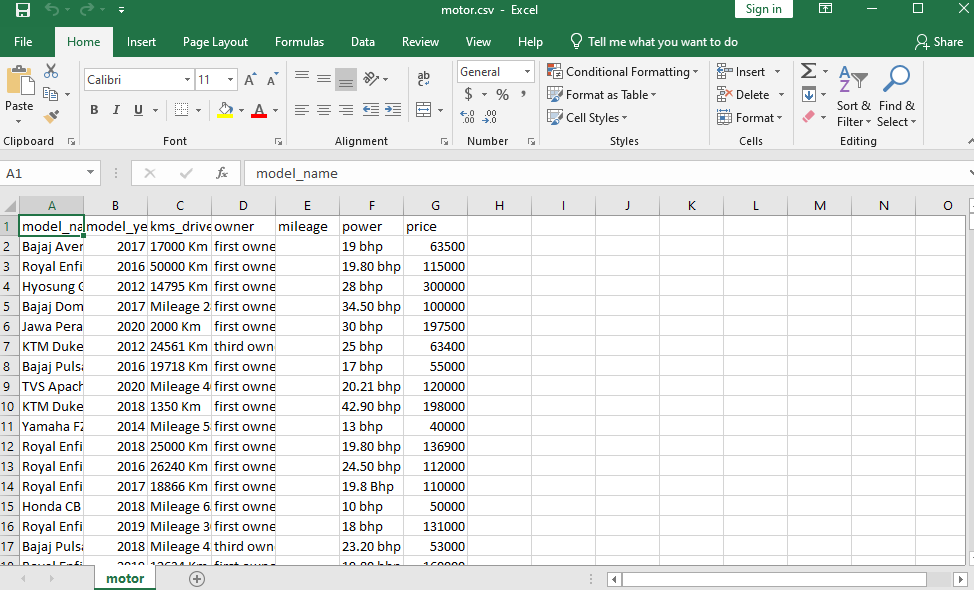
Hồi quy tuyến tính được sử dụng để tìm hiểu mối quan hệ tuyến tính giữa các đặc trưng của xe và giá, từ đó mô hình có thể dự đoán giá của các mẫu mới dựa trên các đặc trưng đó.

# **CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

## **2.1. Xử lý data**

### *2.1.1.Thu thập dữ liệu*

Dữ liệu em thu thập từ trên trang web https://www.kaggle.com/datasets.



*Hình 2: Dữ liệu thu thập được*

### *2.1.2.Các data cần xử lý(làm sạch)*

* Xử lý dữ liệu NaN:

Điền giá trị NaN trong cột số liệu: Có thể điền giá trị NaN bằng giá trị trung bình, median hoặc mode của cột đó.

Loại bỏ dữ liệu NaN: Nếu không cần thiết, bạn có thể loại bỏ các dòng hoặc cột chứa giá trị NaN.

* Xử lý chuỗi:

Loại bỏ khoảng trắng dư thừa: Sử dụng hàm strip() để loại bỏ khoảng trắng dư thừa ở đầu và cuối chuỗi.

Chuyển đổi chuỗi thành chữ hoa hoặc chữ thường: Sử dụng các phương thức upper() hoặc lower() để chuyển đổi chuỗi thành chữ hoa hoặc chữ thường.

* Xử lý ngày tháng:

Chuyển đổi chuỗi ngày tháng thành đối tượng datetime: Sử dụng hàm pd.to\_datetime() để chuyển đổi chuỗi ngày tháng thành đối tượng datetime.

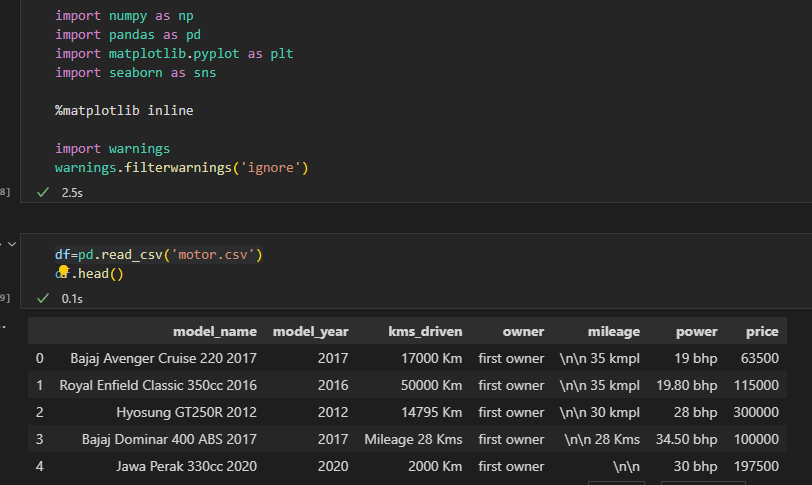
Trích xuất thông tin ngày, tháng, năm: Sử dụng các phương thức như dt.year, dt.month, dt.day để trích xuất thông tin ngày, tháng và năm từ cột ngày tháng.

* Lọc nhiễu:

Loại bỏ dòng trùng lặp: Sử dụng hàm drop\_duplicates() để loại bỏ các dòng trùng lặp trong dữ liệu.

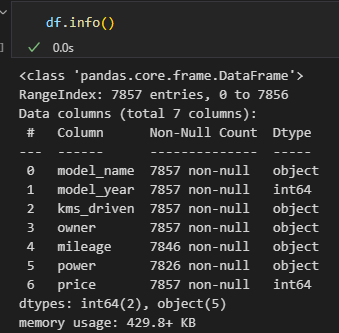
Xử lý ngoại lệ: Có thể áp dụng các phương pháp thống kê như phân phối chuẩn hoặc giá trị ngưỡng để loại bỏ các giá trị nhiễu ngoại lệ.

### *2.1.3.Xử lý data với python*



*Hình 3 : Data trên python*

Đầu tiên sử dụng thư viện pandas để đọc file csv. Sau khi xem qua dữ liệu đầu vào, sẽ xem các thông số thuộc tính,… để từ đó đánh giá dữ liệu đầu vào nên làm những gì để phù hợp với mô hình.



*Hình 4: Thông tin dữ liệu*

Ở đây có 2 loại kiểu dữ liệu là object và int64, object là kiểu dữ liệu gốc và là cơ sở cho tất cả các kiểu dữ liệu khác còn int64 là kiểu dữ liệu số nguyên.Trong bảng thông tin gồm có 7 cột chứa mỗi thuộc tính khác nhau.

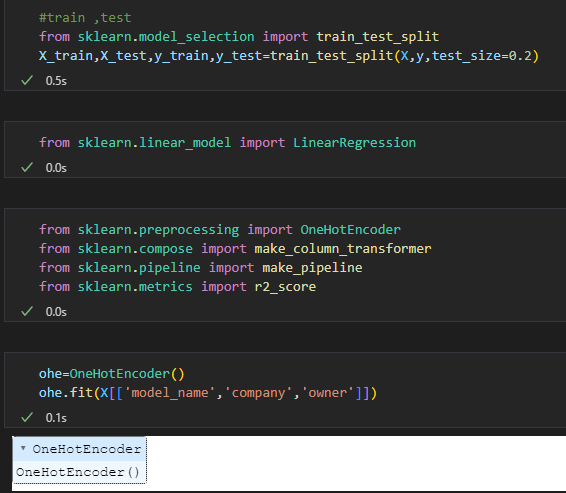
## **2.2. Model**

Mô hình được sử dụng là Hồi quy tuyến tính (Linear Regression). Hồi quy tuyến tính là một phương pháp học máy được sử dụng để dự đoán giá trị liên tục dựa trên một tập hợp các biến độc lập.

Đầu tiên, dữ liệu được chia thành hai tập X và y. Tập X chứa các biến độc lập bao gồm 'company', 'model\_name', 'year', 'kms\_driven' và 'power', trong khi tập y chứa biến phụ thuộc là 'Price' - giá xe.

Tiếp theo, một số công việc tiền xử lý dữ liệu được thực hiện trước khi áp dụng mô hình. Một bước quan trọng là chuyển đổi các biến định danh (categorical variables) trong tập X thành dạng số sử dụng One-Hot Encoding. Điều này đảm bảo rằng các biến định danh có thể được sử dụng trong mô hình hồi quy tuyến tính.

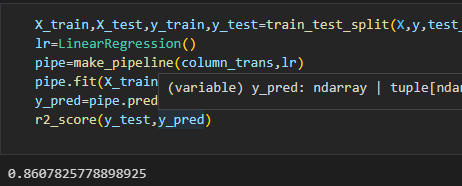
Sau đó, một pipeline (đường ống xử lý dữ liệu) được tạo bằng cách kết hợp các bước xử lý dữ liệu và mô hình hồi quy tuyến tính. Pipeline bao gồm một ColumnTransformer để xử lý các biến định danh và giữ nguyên các biến số, và mô hình hồi quy tuyến tính.



*Hình 5 :Mô hình*

Tiếp theo, mô hình được huấn luyện trên tập huấn luyện (X\_train, y\_train) bằng cách gọi phương thức fit() trên pipeline. Sau đó, mô hình được sử dụng để dự đoán giá trị y trên tập kiểm tra (X\_test) bằng phương thức predict() và độ đo R-squared (R2 score) được tính để đánh giá mức độ phù hợp của mô hình. R2 score là một độ đo thống kê cho biết mức độ phù hợp của mô hình hồi quy tuyến tính với dữ liệu thực tế, nằm trong khoảng từ 0 đến 1. Giá trị càng gần 1 cho thấy mô hình dự đoán tốt hơn.

Cuối cùng, mô hình được lưu vào tệp 'LinearRegressionModel.pkl' bằng cách sử dụng pickle để có thể sử dụng xây dựng web.



*Hình 6 : Kết quả R2*

# **CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH**

## **3.1.Thuật toán**

Tiền xử lý dữ liệu:

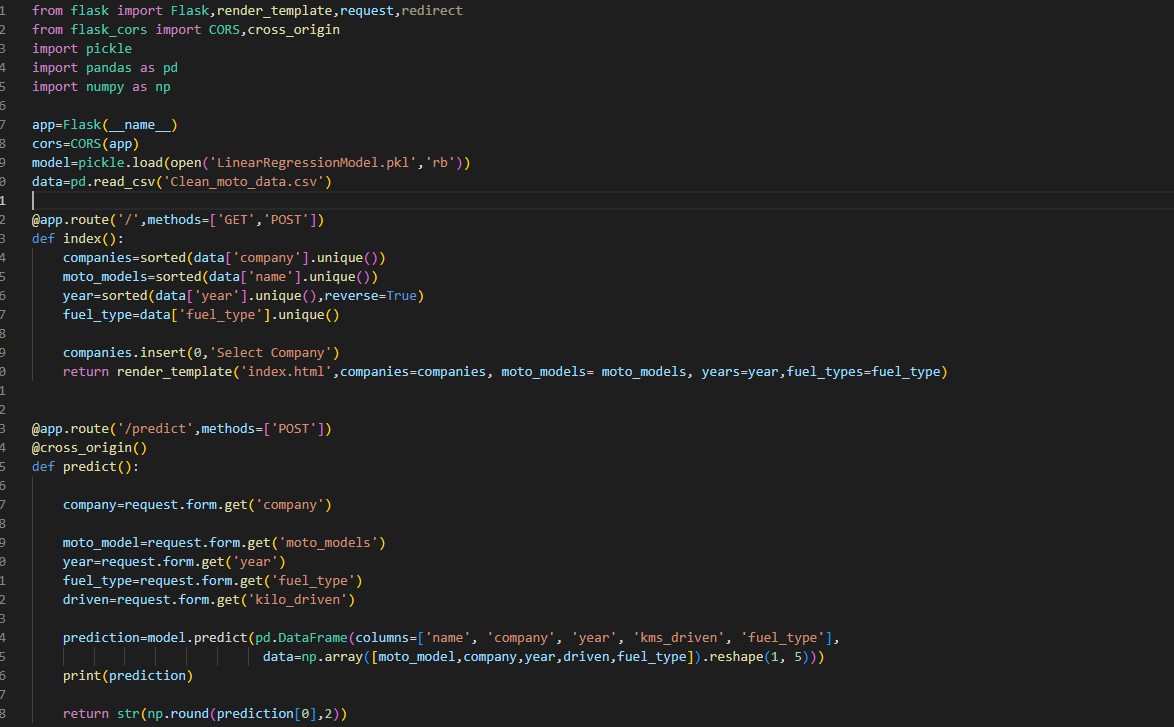
* Đọc dữ liệu từ tệp CSV và loại bỏ các cột không cần thiết.
* Tách cột 'Price' thành hai cột 'price' và 'multiply' để lấy giá tiền và đơn vị đo.
* Chuyển đổi cột 'multiply' thành giá trị số.
* Chuyển đổi cột 'price' sang kiểu dữ liệu số.
* Xử lý các giá trị không hợp lệ và loại bỏ các dòng dữ liệu có giá trị 'Price' bị thiếu.
* Tạo cột 'company' từ đầu của cột 'name'.
* Giới hạn cột 'name' lại thành 3 từ đầu tiên.
* Xóa các cột không cần thiết và loại bỏ các dòng dữ liệu có giá trị thiếu.

Huấn luyện mô hình hồi quy tuyến tính:

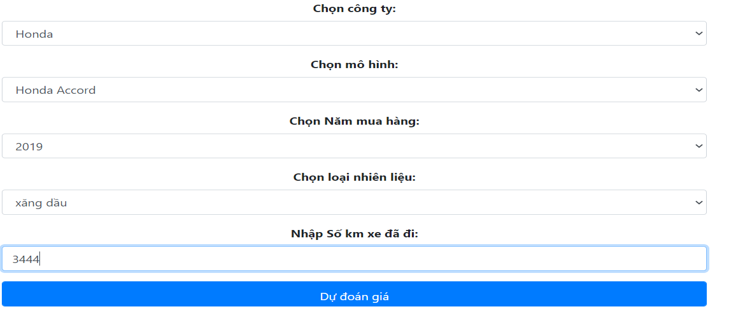
* Mã hóa các biến phân loại bằng OneHotEncoder.
* Xây dựng pipeline để kết hợp việc mã hóa và mô hình hồi quy tuyến tính.
* Chia dữ liệu thành tập huấn luyện và tập kiểm tra.
* Huấn luyện mô hình trên tập huấn luyện.

Dự đoán giá xe trên tập kiểm tra và tính toán hệ số xác định R^2.

## **3.2.Xây dựng chương trình**

****

*Hình 7:Chương trình*



*Hình 8:Giao diện chương trình*

# **CHƯƠNG 4: KẾT LUẬN**

Em đã thực hiện xử lý dữ liệu trên tập dữ liệu xe máy và xây dựng một mô hình dự đoán giá xe dựa trên thông tin như công ty sản xuất, tên xe, năm sản xuất, số kilômét đã đi. Mô hình dự đoán được đánh giá cao với một R-squared (R^2) đạt tới 0.8628, cho thấy mức độ chính xác cao trong việc dự đoán giá trị xe .

Với mô hình này, em tạo ra một trang web dự đoán giá xe máy cho người dùng. Trang web sẽ cho phép người dùng nhập thông tin về công ty sản xuất, tên xe, năm sản xuất, số kilômét đã đi loại xe mà họ quan tâm. Sau khi người dùng nhập thông tin, mô hình sẽ áp dụng dự đoán và hiển thị kết quả giá trị dự đoán của xe máy đó.

Trang web dự đoán giá xe máy sẽ giúp người dùng có cái nhìn trước về giá trị của xe mà họ quan tâm dựa trên các thông tin cơ bản. Điều này có thể hữu ích cho người mua xe máy khi họ đang tìm kiếm một xe mới hoặc muốn biết giá trị hiện tại của xe cũ của mình.

* Ưu điểm :
* Chương trình đã có thể truyền dữ liệu từ web vào model và đưa ra kết quả dự đoán.
* Tốc độ nhanh không bị gián đoạn.
* Giao diện thân thiện dễ sử dụng.
* Nhược điểm :
* Độ sai lệch còn lớn
* Dữ liệu còn ít thuộc tính
* Còn chưa có dữ liệu ở Việt Nam
* Hướng phát triển :
* Mở rộng dữ liệu: Thu thập dữ liệu từ nhiều nguồn khác nhau để mở rộng phạm vi dự đoán giá xe ô tô. Điều này có thể bao gồm việc thu thập thông tin từ các trang web mua bán ô tô, các báo cáo thị trường và dữ liệu từ các nhà sản xuất ô tô. Điều này giúp cung cấp một tập dữ liệu đa dạng và phong phú hơn cho mô hình dự đoán.
* Cải thiện mô hình dự đoán: Thử nghiệm và đánh giá các mô hình dự đoán khác nhau để tìm ra mô hình tốt nhất cho bài toán dự đoán giá xe ô tô. Các mô hình nâng cao, chẳng hạn như mạng neural hồi quy (RNN) hoặc mô hình học sâu (deep learning), có thể được áp dụng để cải thiện độ chính xác dự đoán.

# **TÀI LIỆU THAM KHẢO**

[1].<https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>

[2]. <https://pythonprogramming.net/convolutional-neural-network-kats-vs-dogs-machine-learning-tutorial/>

[3]. <https://machinelearningcoban.com/2016/12/28/linearregression/>