ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH VÀ DỰ ĐOÁN GIÁ XE Ô TÔ

Sinh viên thực hiện:

STT	Họ tên	MSSV	Ngành
1	Phạm Lê Trưởng	20522090	KHMT
2	Trần Phương Thảo	20521938	KHMT

TP. HỒ CHÍ MINH – 12/2023

1. GIỚI THIỆU

Đề tài trong đồ án này là về phân tích thăm dò các đặc trưng ảnh hưởng đến giá xe và sau đó xây dựng mô hình máy học dự đoán giá xe ô tô.

Bộ dữ liệu tự phân tích và tự thu thập tại đường dẫn. Bộ dữ liệu và đề tài do nhóm tự phân tích thiết kế và không dựa trên đề tài nào khác.

Thông qua thực hiện các công cụ, giải pháp thăm dò, thuật toán:

- Thu thập dữ liệu (Selenium, BeautifulSoup).
- Tiền xử lý dữ liệu (Chuẩn hóa, xử lý nan value, xử lý ngoại lệ).
- Phân tích thăm dò (Phân tích bằng các phương pháp thống kê để thăm dò tương quan/ảnh hưởng của các đặc trưng).
- Xây dựng mô hình máy học (Áp dụng các kĩ thuật học máy, xây dựng pipeline, tối ưu hóa mô hình Linear regression).
- Đánh giá hiệu suất (Sử dụng độ đo R2 score để đánh giá mô hình dự đoán so với thực tế).

Kết quả đạt được khá tốt với R2 score là 0.84.

2. MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu về giá xe ô tô.

Bộ dữ liệu tự thu thập tại https://carvago.com/cars

Bộ dữ liệu gồm có:

- 23 cột thuộc tính:
    + Biến số: Seats, Power, Engine capacity, CO2 emissions, Mileage, First Registration, AC charging time, Battery capacity, Warranty Until, Price.
    + Biến phân loại: Make, Model, Body color, Interior color, Interior material, Body, Doors, Fuel, Transmission, Drive type, Emission class, Battery type, Previous owners.
- 1409 dòng dữ liệu.

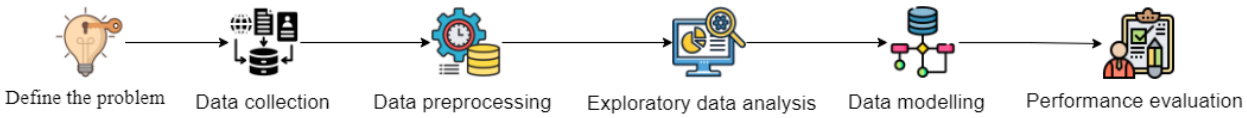
Các thuộc tính của bộ dữ liệu:

Tên thuộc tính	Mô tả thuộc tính	Khoảng giá trị
Make	Hãng sản xuất.	Fiat, Ford, Volkswagen, Audi, BMW, MINI, Skoda, Mercedes-Benz, Seat, Opel, Huyndai, Volvo…
Model	Dòng xe.	500X, Focus, T-Roc, S5, 118, Cooper S Cabrio, Enyaq, Sprinter, Panda, Leon, Kodiaq, A 200, Astra, Ranger …
Body color	Màu sắc thân xe.	Blue, Black, White, Grey, Red, Silver, Green, Yellow, Brown, Orange, Begie.
Interior color	Màu sắc nội thất.	Other interior color, Grey interior, Black interior, Brown interior, Begie interior.
Interior material	Chất liệu nội thất.	Full leather interior, Cloth interior, Alcantara interior, Other interior material, Part leather interior, Velour interior.
Body	Kiểu dáng.	SUV / offroad, Station Wagon, Cabriolet, Coupe, Hatchback, Cargo VAN, Sedans / saloons, Pick-up, MPV, MPV/VAN.
Doors	Số cửa.	4/5 doors, 2/3 doors.
Seats	Số ghế.	1, 2, 3, 4, 5, 6, 7, 8+, 9+.
Fuel	Loại nhiên liệu.	Petrol, Diesel, Electric, Hubrid, CNG, LPG, Other fuel type.

Transmission	Hộp số.	Automatic, Manual
Drive type	Cầu xe.	4x2, 4x4
Power	Mã lực.	Khoảng giá trị từ 44 đến 450.
Engine capacity	Công suất động cơ.	Khoảng giá trị từ 1 đến 5461.
CO2 emissions	Lượng khí thải CO2.	Khoảng giá trị từ 0 đến 288.
Emission class	Loại khí thải.	Euro 6d, No emission class, Euro 6d-TEMP, Euro 6, Euro 5, Euro 6c.
Mileage	Số kilometer xe đã chạy.	Khoảng giá trị từ 0 đến 174500.
First registration	Năm đăng ký xe.	2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022.
Battery type	Loại ắc quy.	No battery class, Lithium-ion (Li-on), Lithium polymer (Li-pol).
AC charging time	Thời gian sạc.	Khoảng giá trị từ 0 đến 12.
Battery capacity	Dung lượng ắc quy.	Khoảng giá trị từ 0 đến 95.
Previous owners	Số người sở hữu xe trước đó.	Unknow, 1, 2, 3, 4.
Warranty until	Thời gian bảo hành.	Khoảng giá trị từ 1 đến 9.
Price	Giá xe.	Khoảng giá trị từ 4699 đến 73299.

3. PHƯƠNG PHÁP PHÂN TÍCH

Quy trình phân tích dữ liệu:



Hình 1: Tổng quan quy trình phân tích dữ liệu.

Define the problem (Định nghĩa vấn đề): Cần phân tích dữ liệu về giá xe ô tô.

Data collection (Thu thập dữ liệu):

- Tìm kiếm các website có bán ô tô.
- Lựa chọn website đầy đủ các yếu tố cho việc phân tích.
- Công cụ thu thập dữ liệu: Selenium, BeautifulSoup.

Data preprocessing (Tiền xử lý dữ liệu):

- Kiểm tra số lượng dữ liệu trống.
- Loại bỏ dữ liệu trùng lặp.
- Chuẩn hóa và thay đổi kiểu dữ liệu phù hợp cho từng đặc trưng.
- Điền khuyết giá trị:
    + Thực hiện điền giá trị Mean cho các khuyết dạng số.
    + Thực hiện điền giá trị Mode cho các khuyết dạng phân loại.
    + Phát hiện đặc trưng Make, Model, Fuel, CO2 emissions, Power không khuyết trong dữ liệu.
    + Giá trị điền khuyết phải dựa trên các yếu tố ràng buộc:
        ▪ Dựa vào Make, Model để điền khuyết: Body color, Interior color, Interior material, Doors, Seats, Engine capacity.
        ▪ Dựa vào Fuel, CO2 emissions để điền khuyết: Emission class, Battery type.
        ▪ Dựa vào Power, Battery type để điền khuyết: AC charging time, Battery capacity.
        ▪ Đối với đặc trưng khó lường như ‘Previous owners’ các giá trị khuyết sẽ là ‘Unknown’.

        - Đối với ‘Warranty until’ (tháng/năm kết thúc bảo hành) thì sẽ quy đổi về số năm bảo hành dựa trên ‘First registration’ (năm đăng kí). Các giá trị khuyết dựa theo số năm bảo hành trung bình của Make (hãng sản xuất).
    - Các giá trị vẫn còn bị khuyết (NaN) tiến hành loại bỏ.
    - Các giá trị ngoại lệ (Outlier) tiến hành loại bỏ

Exploratory data analysis (Phân tích thăm dò dữ liệu):

- Kiểm tra dữ liệu của các biến có phù hợp?
- Bộ dữ liệu có bị khuyết giá trị?
- Phân loại các biến số trong bộ dữ liệu và cho biết biến nào tương đối gần mức đối xứng, biến nào đang mất cân đối.
- Trong tất cả các biến kiểu số thì biến nào có ảnh hưởng đến giá xe.
- Trong các biến kiểu số của bộ dữ liệu có khả năng ảnh hưởng đến giá xe thì biến nào ảnh hưởng ít nhất, biến nào ảnh hưởng nhiều nhất.
- Tìm kiểu dáng của xe có khả năng ít ảnh hưởng đến giá nhất.
- Xe thuộc nhà sản xuất nào, kiểu dáng gì sẽ ảnh hưởng đến giá xe nhiều nhất.
- Tìm tập hợp gồm 3 giá trị biến phân loại bất kỳ có khả năng ảnh hưởng đến giá xe nhiều nhất.
- Tìm tập hợp gồm 4 giá trị biến phân loại bất kỳ có khả năng ít ảnh hưởng đến giá xe nhiều nhất.
- Trong tất cả các biến kiểu phân loại của bộ dữ liệu thì biến nào có khả năng ảnh hưởng đến giá xe nhiều nhất.
- Trực quan hóa sử dụng: Power BI, Matplotlib, Seaborn.

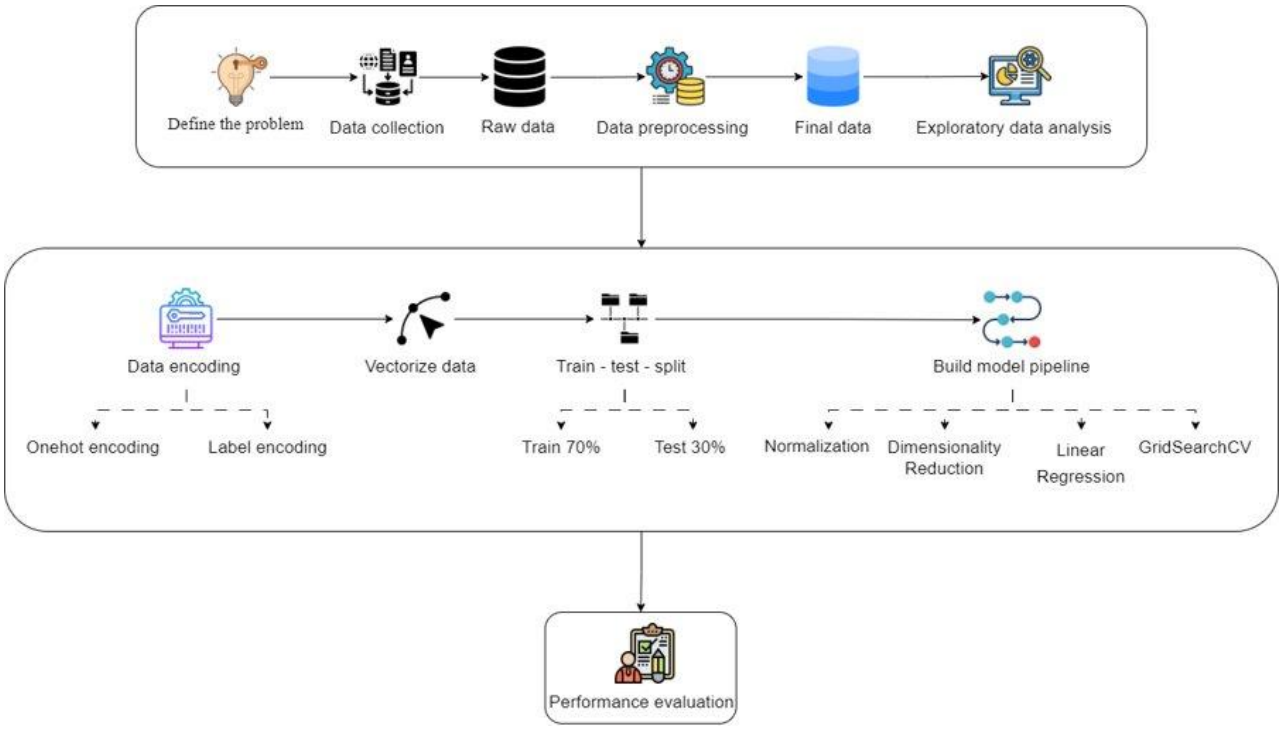
Data modelling (Mô hình hóa dữ liệu):

- Data encoding (Mã hóa dữ liệu):
    + Mục tiêu chuyển đổi dữ liệu dạng phân loại thành dữ liệu số.
    + Phương pháp sử dụng: Onehot Encoding, Label Encoding

- Vectorize data (Vector hóa dữ liệu): Chuyển dữ liệu sau khi mã hóa thành ma trận các vector đặc trưng.
- Train-test-split (Chia tập huấn luyện-kiểm tra):
    + Tập huấn luyện chiếm 70% tổng số dữ liệu: mục tiêu để huyến luyện mô hình.
    + Tập kiểm tra chiếm 30% tổng số dữ liệu: mục tiêu để kiểm thử tính ứng dụng mô hình vào thực tế.
- Build model Pipeline (Xây dựng đường ống dữ liệu) theo các bước:
    + Normalization (Chuẩn hóa dữ liệu): Min-Max Normalization.
    + Dimensionality Reduction(Giảm chiều giữ liệu): PCA algorithm.
    + Mô hình: Linear Regression.
    + Kết hợp Pipeline với GridSearchCV để tìm bộ tham số tốt nhất cho mô hình

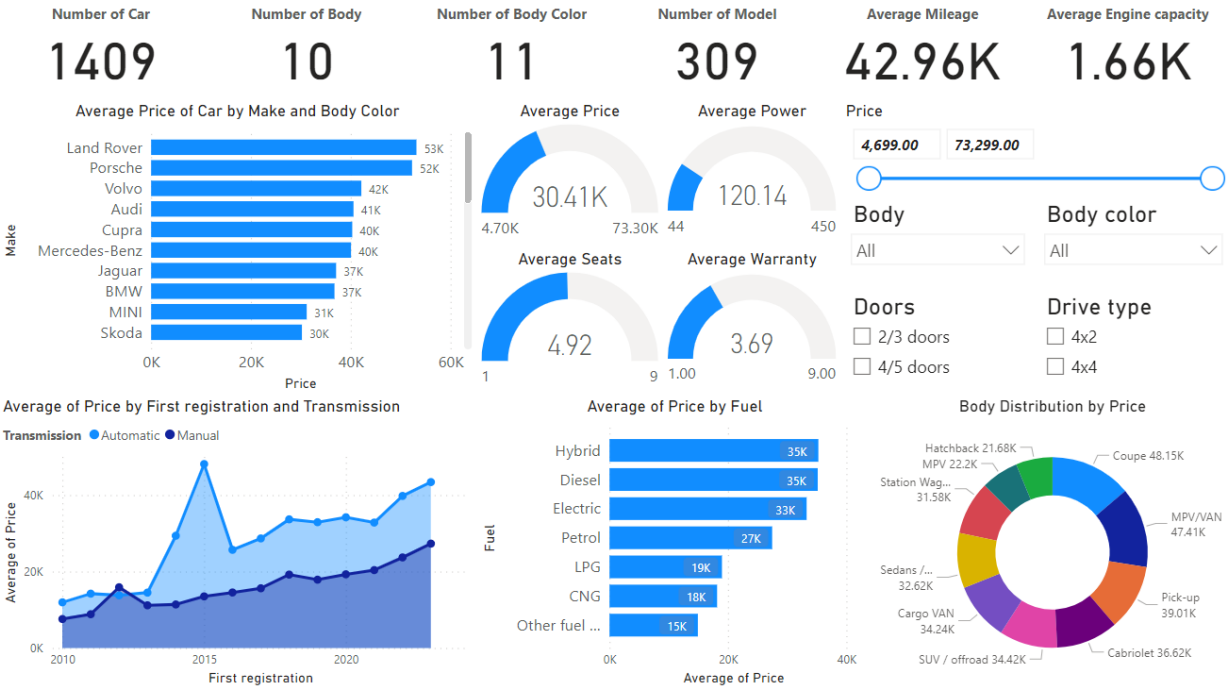
Performance evaluation (Đánh giá hiệu suất):

- Sử dụng mô hình tốt nhất sau khi huấn luyện trên tập train.
- Dự đoán kết quả trên tập test.
- Đánh giá kết quả dự đoán so với thực tế bằng độ đo R2 score



Hình 2: Chi tiết các bước thực hiện cả quy trình.

4. PHÂN TÍCH THĂM DÒ



Hình 3: Dashboard dữ liệu.

Các phân tích thăm dò đã kiểm thử:

- Hai hãng xe có giá trung bình cao nhất là: Land Rover và Porsche.
- Giá thành các xe có hộp số tự động (Automatic) cao hơn hộp số thủ công(Manual) qua các năm và điều này cũng đúng cho các hãng khác nhau.
- Nguồn nhiên liệu hầu hết được ưa chuộng là: Hybrid, Diesel, Electric.
- Kiểu dữ liệu các đặc trưng đã phù hợp.
- Bộ dữ liệu đã không còn khuyết giá trị.
- Danh sách biến số gần mức đối xứng: Power, First registration.
- Danh sách biến số mất cân đối: Engine capacity, CO2 emissions, Mileage, AC charging time, Battery capacity, Warranty until.
- Trong các biến kiểu số thì biến ảnh hưởng đến giá xe là: Power, Engine capacity, Mileage, First registration.
- Biến số có nhiều ảnh hưởng đến giá xe nhất là: Power.
- Biến số có ít ảnh hưởng đến giá xe nhất là: Mileage.
- Kiểu dáng xe có ít ảnh hưởng đến giá xe nhất là: Pick-up.

- Xe thuộc nhà sản xuất 'Hyundai' kiểu dáng 'MPV/VAN' ảnh hưởng đến giá xe nhất.
- Tập hợp 3 biến phân loại 'Doors', 'Transmission', 'Drive type' ảnh hưởng đến giá xe nhiều nhất.
- Tập hợp 4 biến phân loại 'Body color', 'Interior color', 'Emission class', 'Previous owners' ít ảnh hưởng đến giá xe nhất.
- Trong tất cả các biến kiểu phân loại thì biến 'Transmission' ảnh hưởng đến giá xe nhiều nhất.

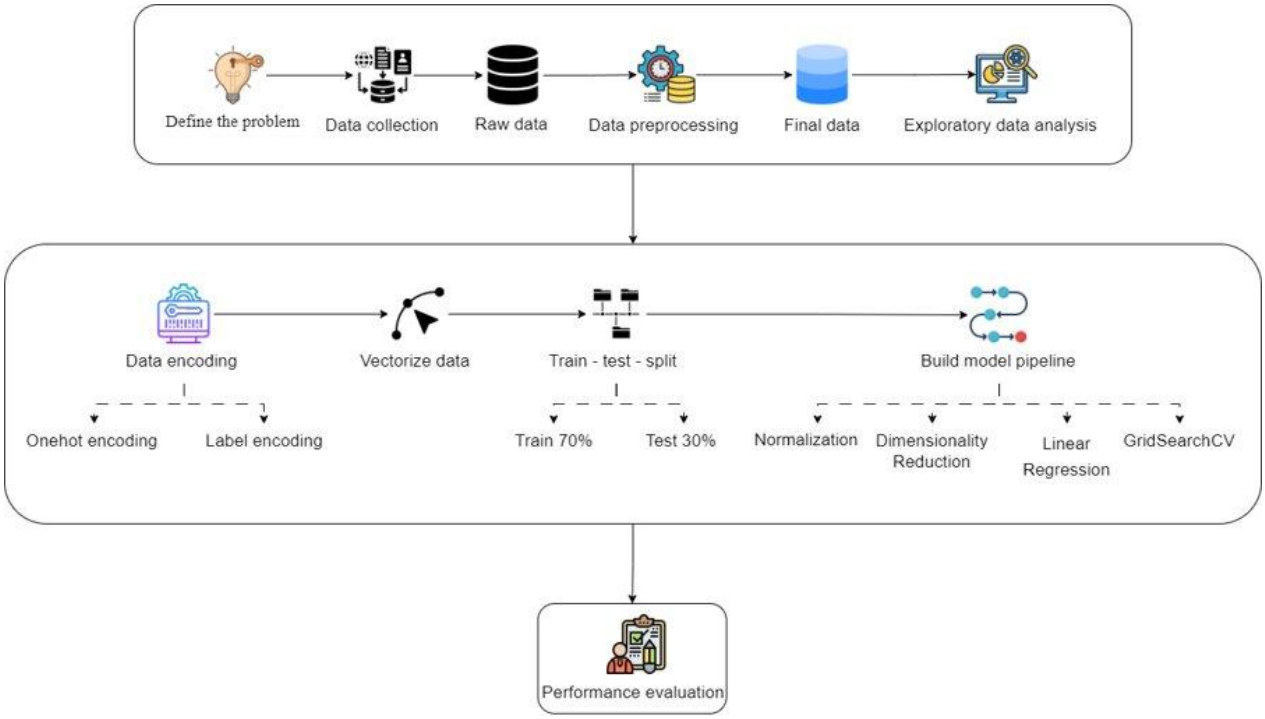
5. KẾT QUẢ PHÂN TÍCH

Kết quả phân tích các biến quan trọng (mức độ ảnh hưởng giảm dần) đã thăm dò được:

Đặc trưng số	**Đặc trưng phân loại**
- Power. - Engine capacity. - First registration. - Mileage.	- Transmission. - Drive type. - Interior material. - Body. - Emission class. - Make. - Fuel. - Model. - Interior color. - Previous owners. - Seats. - Doors. - Body color. - Battery type

6. KẾT LUẬN

Tổng thể chi tiết toàn bộ quá trình:



Kết quả thu được trên tập thử nghiệm R2 score = 0.84

TÀI LIỆU THAM KHẢO

Scikit-learn. https://scikit-learn.org/stable/ (Truy cập 5/12/2023)

Geeksforgeeks. https://www.geeksforgeeks.org/ (Truy cập 5/12/2023)

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Phạm Lê Trưởng	- Thu thập dữ liệu. - Tiền xử lý dữ liệu. - Phân tích thăm dò dữ liệu. - Tạo dashboard dữ liệu. - Mô hình hóa dữ liệu. - Đánh giá hiệu suất.
2	Trần Phương Thảo	- Giới thiệu đề tài. - Tìm kiếm website phù hợp cho đề tài - Thống kê dữ liệu - Mô tả bộ dữ liệu. - Thiết kế và chỉnh sửa template Word, PDF, Powerpoint.