ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

MÔN HỌC: LẬP TRÌNH PYTHON CHO MÁY HỌC

ĐỀ TÀI: Car Price Prediction

Giảng viên hướng dẫn	**:**	**TS.Nguyễn Vinh Tiệp**
Sinh viên thực hiện	**:**	**Nguyễn Võ Ngọc Bảo - 23520131**
		Ngô Phương Nam - 23520974
		Vũ Việt Cương - 23520213
Lớp	**:**	**CS116.P21**

Thành phố Hồ Chí Minh, ngày 27 tháng 03 năm 2025

1. Giới thiệu đề tài

A. Tên đề tài:

Car Price Prediction

B. Mô tả nhiệm vụ:

1. Dữ liệu hiện có:

- Tập dữ liệu huấn luyện (train.csv) bao gồm các thông tin đã biết về
  + Đặc điểm kỹ thuật: hãng xe (model), năm sản xuất (year), loại động cơ (motor_type), dung tích động cơ (motor_volume), kiểu xe (type), màu sắc (color), hướng vô lăng (wheel).
  + Tình trạng xe: Số km/miles đã chạy (running), trạng thái (status như "excellent", "good", "crashed").
  + Giá cả: Giá bán (price) - đây là trường mục tiêu (target variable) trong bài toán dự đoán
- Tập dữ liệu kiểm tra (test.csv) bao gồm:
  + ID: Số định danh duy nhất cho từng mẫu dữ liệu.
  + Đặc điểm kỹ thuật: hãng xe (model), năm sản xuất (year), loại động cơ (motor_type), dung tích động cơ (motor_volume), kiểu xe (type), màu sắc (color), hướng vô lăng (wheel).
  + Tình trạng xe: Tình trạng xe: Số km/miles đã chạy (running), trạng thái (status như "excellent", "good", "normal").

Khác với tập train, tập test không có cột price vì đây là trường mục tiêu cần dự đoán

- File mẫu nộp bài (sample_submission.csv) thể hiện định dạng yêu cầu khi gửi kết quả dự đoán

2. Mục Tiêu:

- Dự đoán giá trong tập dữ liệu kiểm tra (test.csv) chính xác nhất có thể, sử dụng chỉ số sMAPE (Symmetric Mean Absolute Percentage Error) làm thước đo sai số.

2. Một số số liệu ban đầu về quy mô dữ liệu

C. File Train.csv

- 1643 dòng
- 10 cột:

Tên Trường	Kiểu dữ liệu	Ý nghĩa
Model	Chuỗi	Hãng xe

Year	Số Nguyên	Năm sản xuất
Motor_Type	Chuỗi	Loại động cơ
Running	Chuỗi	Số km/miles đã đi
Wheel	Chuỗi	Vị trí vô lăng
Color	Chuỗi	Màu xe
Motor_Type	Chuỗi	Kiểu xe
Status	Chuỗi	Tình trạng xe
Motor_Volume	Số thực	Dung tích động cơ
Price	Số nguyên	Giá xe (mục tiêu)

D. File Test.csv

- 412 dòng
- 10 cột:

Tên Trường	Kiểu dữ liệu	Ý nghĩa
ID	Số Nguyên	Định danh từng dòng
Model	Chuỗi	Hãng xe
Year	Số Nguyên	Năm sản xuất
Motor_Type	Chuỗi	Loại động cơ
Running	Chuỗi	Số km/miles đã đi
Wheel	Chuỗi	Vị trí vô lăng
Color	Chuỗi	Màu xe
Motor_Type	Chuỗi	Kiểu xe
Status	Chuỗi	Tình trạng xe

Motor_Volume	Số thực	Dung tích động cơ

3. Input và Output

E. Input

Dữ liệu Đầu Vào (Input) trong file train.csv và test.csv

- **Các đặc trưng của xe:**
  - Dữ liệu đầu vào chứa các cột mô tả đặc điểm của từng chiếc xe, chẳng hạn như:
    - **ID:** Mã định danh duy nhất cho mỗi mẫu xe.
    - **Hãng xe và Model:** Thông tin về nhãn hiệu, dòng xe (các thông tin này giúp xác định mức giá theo thương hiệu và mẫu xe).
    - **Năm sản xuất:** Thời điểm sản xuất, ảnh hưởng lớn đến giá trị của xe.
    - **Số km đã đi:** Một chỉ số quan trọng phản ánh mức độ sử dụng của xe.
    - **Các đặc điểm kỹ thuật:** Có thể bao gồm dung tích động cơ, loại động cơ, hệ thống truyền động, hộp số, và các thông số kỹ thuật khác.
    - **Các thông tin bổ sung:** Những đặc trưng khác có thể có như màu sắc, tình trạng xe (mới hay đã qua sử dụng), số cửa,…
- **Quá trình xử lý dữ liệu:**
  - **Tiền xử lý:** Các đặc trưng ban đầu thường cần được làm sạch dữ liệu, xử lý giá trị thiếu, và chuyển đổi dữ liệu dạng văn bản thành dạng số (ví dụ: one-hot encoding đối với dữ liệu phân loại).
  - **Chuẩn hóa/biến đổi:** Các đặc trưng số có thể cần được chuẩn hóa (scaling) để phù hợp với thuật toán học máy.
- **Mục đích:**
  - Các dữ liệu input này sẽ được sử dụng để huấn luyện mô hình dự đoán giá xe.
  - Trong quá trình huấn luyện, ma trận các đặc trưng (X) được trích xuất từ dữ liệu đầu vào, và cột giá (price) trong file train.csv sẽ làm giá trị mục tiêu.

F. Output

Kết quả Dự Đoán (Output) trong file submission

- **Cấu trúc file output (sample_submission.csv):**
  - **ID:** Cột đầu tiên chứa mã định danh tương ứng với mỗi mẫu xe trong tập test.
  - **Price:**
    - Cột thứ hai chứa giá trị dự đoán (giá xe) cho từng mẫu.
    - Giá trị dự đoán này thường được tính toán bằng mô hình học máy sau khi được huấn luyện trên dữ liệu input.
- **Yêu cầu định dạng:**
  - Số lượng dòng trong file output phải trùng khớp với số dòng trong test.csv.
  - Tên cột và định dạng dữ liệu phải đúng theo mẫu mẫu (template) đã cho bởi cuộc thi trên Kaggle.
  - Các giá trị dự đoán nên được biểu diễn dưới dạng số thực (float) với độ chính xác hợp lý, phù hợp với yêu cầu của cuộc thi.
- **Mục đích:**
  - File output (F) là kết quả cuối cùng của đồ án, dùng để so sánh với giá trị thực tế của xe (trong trường hợp có sẵn thông tin hay trong quá trình chấm điểm tự động của cuộc thi Kaggle).
  - File này sẽ được nộp lên hệ thống đánh giá của Kaggle để xác định hiệu quả của mô hình dự đoán.