

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA TOÁN – TIN HỌC**



## **BÁO CÁO BÀI TẬP LỚN**

### **Đề tài: Dự đoán giá xe ô tô**

**Môn: Python cho khoa học dữ liệu**

Giảng viên: T.S Nguyễn Tấn Trung

Nhóm: TNT

Tp. Hồ Chí Minh, 16/01/2022

## **Mục lục**

<b>I.</b>	<b>Giới thiệu .....</b>	<b>3</b>
1.	Thông tin nhóm: .....	3
2.	Giới thiệu đề tài: .....	3
3.	Hướng giải quyết đề tài: .....	3
4.	Bảng phân công công việc: .....	3
<b>II.</b>	<b>Chuẩn bị dữ liệu .....</b>	<b>5</b>
1.	Thu thập dữ liệu: .....	5
2.	Chuyển đổi dữ liệu thô: .....	5
3.	Tiền xử lý dữ liệu: .....	5
<b>III.</b>	<b>Xây dựng mô hình.....</b>	<b>6</b>
1.	Quá trình: .....	6
2.	Mô hình: .....	6
<b>IV.</b>	<b>Trực quan hoá dữ liệu .....</b>	<b>6</b>
1.	Quá trình: .....	6
2.	Các biểu đồ: .....	6
<b>V.</b>	<b>Tài liệu tham khảo .....</b>	<b>7</b>

## I. Giới thiệu

### 1. Thông tin nhóm:

Tên nhóm: TNT

STT	Họ tên	MSSV	Email	Ghi chú
1	Đoàn Quang Nhật Tài	19110431	19110431@student.hcmus.edu.vn	Nhóm trưởng
2	Trần Quang Nghĩa	19110392	19110392@student.hcmus.edu.vn	
3	Huỳnh Thị Bảo Trân	19110482	19110482@student.hcmus.edu.vn	

### 2. Giới thiệu đề tài:

Dự đoán giá xe ô tô.

### 3. Hướng giải quyết đề tài:

Từ câu hỏi được đặt ra: “Dự đoán về giá xe dựa trên các đặc tính nào của nó?”. Với dữ liệu đầu vào là các đặc trưng, tính chất của xe. Nhóm em tiến hành chuẩn bị, phân tích, xử lý dữ liệu để đưa ra kết quả đầu ra là giá xe dự đoán.

Đầu tiên, về việc chuẩn bị dữ liệu, nhóm em tiến hành thu thập dữ liệu. Sau khi crawl dữ liệu, nhóm đã đọc tìm hiểu các thuộc tính. Về sơ bộ, dữ liệu có khoảng 15 dòng xe với hơn 200 thuộc tính liên quan đến từng đặc trưng của xe. Từ đó, tiến hành làm sạch, chuyển đổi tập dữ liệu thô, tiền xử lý dữ liệu (chia các thuộc tính, loại bỏ dữ liệu nhiễu...) sao cho phù hợp bài toán.

Thứ hai, nhóm đã tiến hành xây dựng các mô hình hồi quy tuyến tính để tính toán và xử lý dữ liệu, đưa ra các dự đoán về giá xe. Chi tiết cụ thể, nhóm em sẽ trình bày ở trong phần sau.

Cuối cùng, để làm dữ liệu trở nên dễ hiểu và sinh động, nhóm em đã tiến hành trực quan hoá các dữ liệu thông qua từng loại biểu đồ. Đồng thời đưa ra nhận xét về dữ liệu và kết quả dự đoán giá xe ô tô.

### 4. Bảng phân công công việc:

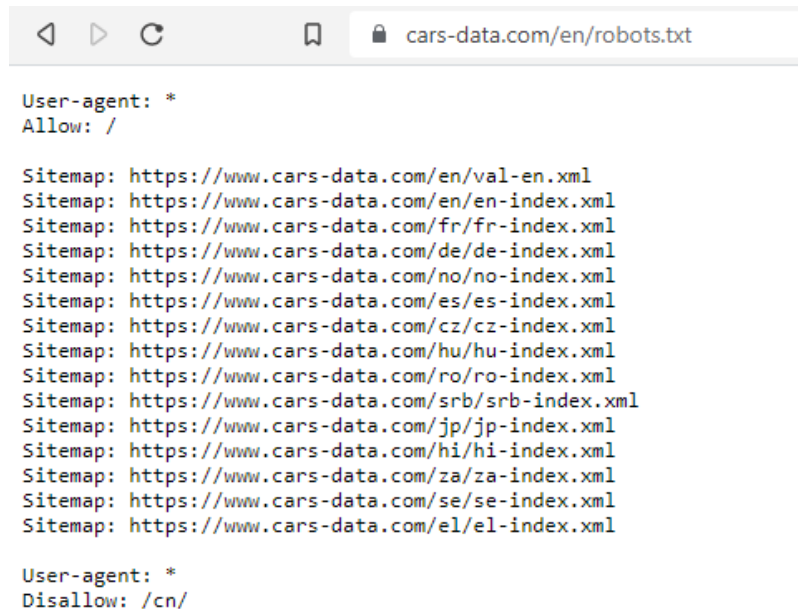
Thời gian	Người thực hiện	Công việc	Kết quả
20/12/2021 (Chọn đề tài)	Nhóm	Thảo luận và chọn đề tài.	Chọn đề tài: “Dự đoán giá xe ô tô”.
27/12/2021 (Làm rõ đề tài)	Nhóm	Phân tích các yếu tố xác định, thực hiện bài toán.	Thực hiện crawl dữ liệu bằng Requests, BeautifulSoup. Xây dựng mô hình hồi quy tuyến tính để dự báo. Trực quan dữ liệu bằng biểu đồ.
06/01/2022- 15/01/2022 (Thực hiện đồ án)	Tài	Chuẩn bị dữ liệu: thu thập dữ liệu, chuyển đổi dữ liệu thô, tiền xử lý các dữ liệu.	Crawl được dữ liệu. Đưa ra được file dữ liệu đã xử lý thô.
	Nghĩa	Xây dựng mô hình hồi quy tuyến tính để dự đoán.	Xây dựng được mô hình hồi quy cơ bản.
	Trân	Trực quan hoá dữ liệu trên bằng các loại biểu đồ.	Trực quan được dữ liệu trên một vài biểu đồ.
16/01/2022 (Đánh giá lần 1)	Nhóm	Kiểm tra tiến độ, chuẩn bị demo. Hiệu chỉnh các sai sót ở từng phần thực hiện. Viết báo cáo.	Mức độ hoàn tất là 65% so với lúc đầu đề ra.
17/01/2022 (Đánh giá lần 2)	Nhóm	Kiểm tra và cải thiện các lỗi ở đánh giá lần 1. Kết nối các phần lại với nhau.	
20/01/2022 (Chuẩn bị báo cáo)	Nhóm	Hoàn thành báo cáo quá trình và slide báo cáo. Kiểm tra lần cuối các kết quả model và mức độ hoàn thành.	

## II. Chuẩn bị dữ liệu

### 1. Thu thập dữ liệu:

Dữ liệu thu thập trên trang: <http://www.cars-data.com/en/>.

Dữ liệu được thu thập hợp pháp và chính thống trên trang web.



```

User-agent: *
Allow: /

Sitemap: https://www.cars-data.com/en/val-en.xml
Sitemap: https://www.cars-data.com/en/en-index.xml
Sitemap: https://www.cars-data.com/fr/fr-index.xml
Sitemap: https://www.cars-data.com/de/de-index.xml
Sitemap: https://www.cars-data.com/no/no-index.xml
Sitemap: https://www.cars-data.com/es/es-index.xml
Sitemap: https://www.cars-data.com/cz/cz-index.xml
Sitemap: https://www.cars-data.com/hu/hu-index.xml
Sitemap: https://www.cars-data.com/ro/ro-index.xml
Sitemap: https://www.cars-data.com/srb/srb-index.xml
Sitemap: https://www.cars-data.com/jp/jp-index.xml
Sitemap: https://www.cars-data.com/hi/hi-index.xml
Sitemap: https://www.cars-data.com/za/za-index.xml
Sitemap: https://www.cars-data.com/se/se-index.xml
Sitemap: https://www.cars-data.com/el/el-index.xml

User-agent: *
Disallow: /cn/
```

Dữ liệu có khoảng hơn 16.000 dòng (Dữ liệu bao gồm khoảng 35 thuộc tính liên quan đến các đặc điểm của ô tô như phụ tùng, trọng lượng, hãng...).

Xây dựng các hàm lấy thuộc tính như: `get_attributes` và `get_car`. Hàm `get_attributes` dùng để truy cập vào từng lớp của trang web để trích xuất các thông tin/dữ liệu cần thiết, hàm `get_car` thu thập các thông tin/dữ liệu theo từng thuộc tính của từng xe trong hãng.

Trang web có 94 hãng, nhưng để tiết kiệm thời gian, nhóm em tiến hành thu thập dữ liệu của 15 hãng đầu tiên để thuận tiện trong việc trực quan và xây dựng các model.

### 2. Chuyển đổi dữ liệu thô:

Chuyển dữ liệu sau khi crawl được thành dữ liệu có thể dùng được theo yêu cầu của bài toán (sẽ cập nhật cụ thể sau).

### 3. Tiền xử lý dữ liệu:

Xử lý dữ liệu có thể dùng được theo yêu cầu của bài toán (sẽ cập nhật cụ thể sau).

### III. Xây dựng mô hình

#### 1. Quá trình:

**Xây dựng thuộc tính mới:** Ta cần thêm một thuộc tính “Tiết kiệm nhiên liệu”, ký hiệu là *fuelconomy*, để đưa ra các thông số cụ thể và đánh giá được sự tiết kiệm nhiên liệu của từng loại xe.

Ta xây dựng dựa trên công thức sau:

$$\text{fuelconomy} = 0.55 * \text{citympg} + 0.45 * \text{hightwaympg}$$

Sau đó ta phân loại công ty dựa trên giá xe của từng công ty như sau:

- Budget: 0 – 10000
- Medium: 10000 – 20000
- Highend: 20000 – 40000

**Phân tích lưỡng biến:** Dựa trên các thông tin có được ở bước trên ta tiếp tục tiến hành phân tích để tìm ra mối liên hệ giữa các biến và suy ra danh sách các biến quan trọng.

**Biến giả:** Định nghĩa hàm dummies; Áp dụng hàm trên cho *cars\_lr*.

#### 2. Mô hình:

Cập nhật sau.

### IV. Trực quan hoá dữ liệu

#### 1. Quá trình:

Đọc hiểu các thuộc tính và số liệu sau khi đã phân tích và tiến hành trực quan, trình bày các dữ liệu thành các biểu đồ. Đồng thời đưa ra các nhận xét về dữ liệu như thế nào (cụ thể sẽ được cập nhật sau).

#### 2. Các biểu đồ:

Biểu đồ boxplot: biểu hiện các dữ liệu ngoại lai theo từng thuộc tính.

Biểu đồ cột: biểu hiện từng hãng xe với mức trung bình của giá xe.

Biểu đồ distplot: biểu hiện mức phân bố của các thuộc tính trong dữ liệu.

## **V. Tài liệu tham khảo**

Cập nhật sau.