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

DỰ ĐOÁN GIÁ ĐỒNG HỒ ĐEO TAY

Sinh viên thực hiện:		
STT	Họ tên	MSSV
1	Phan Quốc Vỹ	21522814
2	Dương Tấn Hoàng	21520866
3	Võ Viết Thuận	21521504
4	Tăng Minh Hiển	21520229

TP. HỒ CHÍ MINH – 11/2023

1. GIỚI THIỆU

- Đồng hồ đeo tay vừa là một chiếc đồng hồ, vừa là một phụ kiện thời trang mà nhiều người ưa thích sử dụng. Ngoài công dụng chính là để xem giờ, đồng hồ đeo tay còn là một món đồ trang sức với nhiều mẫu mã và kiểu dáng khác nhau. Một chiếc đồng hồ khi được cho ra thị trường tùy vào mục đích sử dụng sẽ có những thiết kế khác nhau, và giá thành khác nhau. Giá thành của một chiếc đồng hồ phụ thuộc vào rất nhiều yếu tố như về nhà sản xuất, thông số kỹ thuật, loại hình hiển thị, … và một số yếu tố bên ngoài như thị trường, xu hướng thời trang.

  Đề tài này tập trung vào việc phân tích giá của đồng hồ đeo tay dựa trên các đặc trưng thường có của một chiếc đồng hồ thông qua bộ dữ liệu tự thu thập. Nhóm áp dụng các phương phân tích dữ liệu để xác định các yếu tố ảnh hưởng sau đó áp dụng thuật toán học máy để xây dựng dựa trên các yếu tố này và đưa ra dự đoán cho giá của đồng hồ đeo tay.

- **Bộ dữ liệu:** Dữ liệu sử dụng cho đề tài này được nhóm **tự thu thập** tại website Amazon [1] sử dụng phương pháp bán tự động.

- **Các công cụ, giải pháp, thuật toán và mô hình được áp dụng**:
  + Công cụ: Crawl dữ liệu: Chrome extension nhóm tự viết
  + Giải pháp: Làm sạch dữ liệu bằng các thuật toán xử lý xâu đơn giản. Điền giá trị khuyết: Thuật toán KNN (KNNImputer của sklearn).
  + Thuật toán áp dụng: Mô hình: Cây quyết định, Hồi quy tuyến tính, KNN, Naïve Bayes.

- **Sơ lược về kết quả đạt được**: Mô hình dự đoán tương đối tốt trên dữ liệu test. Với một số điểm dữ liệu hiếm, mô hình dự đoán cách khá xa giá trị ground-truth.

- **Đảm bảo**: Toàn bộ nội dung đồ án được nhóm tự phân tích, tự xây dựng mô hình và không dựa trên bất kỳ công trình, bài tập có sẵn nào khác. Nhóm cam kết không nhận điểm số nếu giảng viên tìm thấy bài làm tương tự trên Internet và chứng minh được bài làm của nhóm dựa trên bài làm đó.

2. MÔ TẢ BỘ DỮ LIỆU

- **Tóm tắt bộ dữ liệu:** Bộ dữ liệu đã được nhóm thu thập sử dụng phương pháp bán tự động, bao gồm các thông tin của đồng hồ đeo tay được cung cấp trên website như về kiểu dáng, thương hiệu, tính năng, vật liệu, xuất xứ, thông số kỹ thuật và đánh giá từ khách mua hàng

- **Phương pháp thu thập:** Vì trang web Amazon có hệ thống nhận diện người/máy, nên không thể sử dụng phương pháp tự động 100% để thu thập. Thay vì vậy, nhóm đã viết một extension [2] để chèn một đoạn mã javascript vào trang web để nó tự động đọc các element, trích xuất thông tin cần thiết, sau đó tải về dưới dạng csv và đóng tab. Điều kiện duy nhất để extension này chạy là phải có người nhấp vào từng link sản phẩm. Đó là lý do phương pháp này là phương pháp bán tự động.

- **Thống kê bộ dữ liệu:**
  + Gồm 26 biến trong đó có 17 biến phân loại và 9 biến số.
  + Có tổng cộng 300 dòng dữ liệu.

- **Mô tả chi tiết các biến của bộ dữ liệu:**

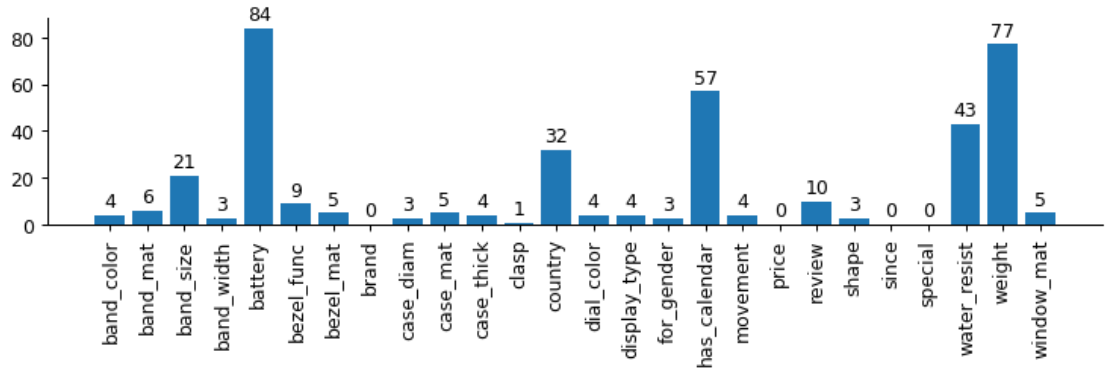
	Attributes	Type	Missing values	Description
1	band_color	object	4	Màu của dây đeo
2	band_mat	object	6	Chất liệu của dây đeo
3	band_size	float32	21	Độ dài dây đeo (đv. cm)
4	band_width	float32	3	Bề rộng dây đeo (đv. cm)
5	battery	object	84	Loại pin
6	bezel_func	object	9	Cách vận hành của bezel
7	bezel_mat	object	5	Chất liệu bezel
8	brand	object	0	Hãng đồng hồ
9	has_calendar	bool	57	Có tính năng lịch hay không

10	case_thick	float32	4	Độ dày của case (đv. cm)
11	case_diam	float32	3	Đường kính của case (cm)
12	case_mat	object	5	Chất liệu của case đồng hồ
13	clasp	object	1	Loại khóa đồng hồ
14	country	object	32	Nước sản xuất
15	since	int32	0	Năm phát hành
16	for_gender	object	3	Phân loại giới tính phù hợp
17	dial_color	object	4	Màu kim
18	window_mat	object	5	Chất liệu mặt đồng hồ
19	display_type	object	4	Loại hiển thị
20	shape	object	3	Hình dạng case đồng hồ
21	weight	float32	77	Khối lượng (đơn vị g)
22	movement	object	4	Động cơ vận hành đồng hồ
23	special	bool	0	Có tính năng đặc biệt khác?
24	water_resist	float32	43	Độ sâu chịu nước (đv. m)
25	review	float32	10	Đánh giá từ khách hàng
26	price	float32	0	Giá sản phẩm

3. TIỀN XỬ LÝ DỮ LIỆU

3.1. Xử lý dữ liệu bị khuyết

- Thống kê số lượng dữ liệu bị khuyết ban đầu:



- Phương pháp xử lý: đối với biến has_calendar, giá trị NaN sẽ được thay bằng False, số còn lại, áp dụng phương pháp điền khuyết KNNImputer với tham số neighbors mặc định bằng 5.
- Lý do chọn phương pháp: Phương pháp xử lý KNN được cho là bảo toàn mối quan hệ giữa các biến với nhau. Bằng cách tìm các điểm dữ liệu đã biết gần với điểm đang được xử lý, sau đó trung bình cộng giá trị của biến cần được điền trên các điểm đó, ta được giá trị phù hợp điền vào vị trí bị khuyết. [3]

3.2. Chuẩn hóa dữ liệu

- Dữ liệu sau khi crawl về sẽ có rất vấn đề khác nhau như thiếu nhất quán giữa các giá trị của biến; giá trị số sử dụng đơn vị khác nhau; các thuộc tính thừa, thuộc tính khóa. Vì thế, ta cần chuẩn hóa dữ liệu và thực hiện một số xử lý:
  + Chuẩn hóa xâu ký tự. Chuyển các ký tự thành dạng thường, chuyển xâu về dạng không dấu, sắp xếp lại các cụm từ trong chuỗi nếu chuỗi là một danh sách các từ để đảm bảo tính thống nhất.
  + Chuyển đổi xâu thành các giá trị số có đơn vị và xử lý các đơn vị đó. Thống nhất các loại đơn vị đo (độ dài: cm; khối lượng: gram, độ sâu: m).
  + Gom các giá trị biến phân loại chỉ xuất hiện 1 lần thành một giá trị chung là ‘other’ để dễ xử lý, tính toán.
  + Loại bỏ các trường không cần thiết như ngày tháng, mã sản phẩm, …
- Dữ liệu kiểm thử được tách riêng trước khi thực hiện thao tác này, để đảm bảo các giá trị khác ‘other’ của các biến phân loại xuất hiện nhiều hơn 1 lần trong dữ liệu huấn luyện.

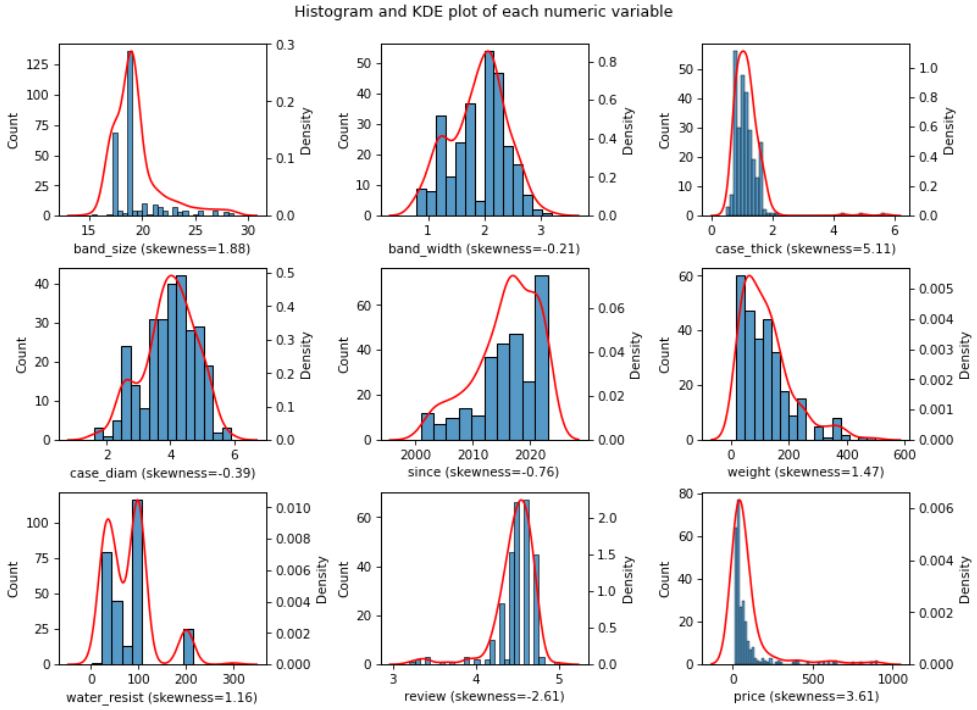
4. PHÂN TÍCH THĂM DÒ

4.1. Phân tích các biến liên tục

- Phân phối và hàm mật độ xác suất của các biến dữ liệu số

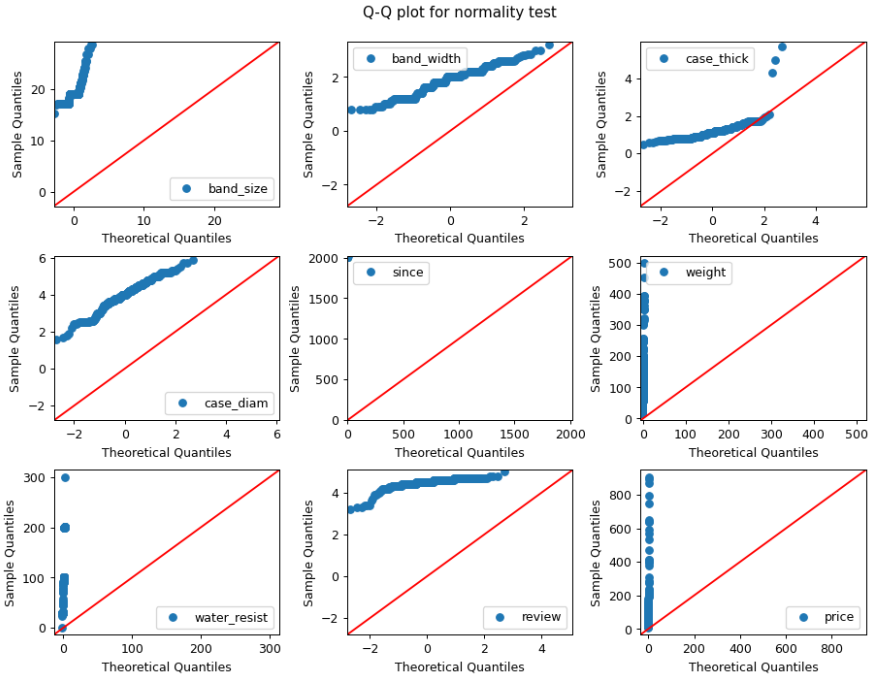
***** *Nhận xét:*

- Nhìn chung, các biến số đều có phân phối tương đối bình thường. Điều này cho thấy rằng các dữ liệu không bị lệch hoặc có hình dạng bất thường khác.
- Các biến số 'band_size', 'case_thick', 'weight', 'water_resists', 'price' có xu hướng tập trung ở các giá trị thấp. Điều này cho thấy các đồng hồ đeo tay trong tập dữ liệu này có kích thước, trọng lượng và đặc biệt là giá thành tương đối thấp.
- Các biến số 'case_diam', 'since', band_width' sự phân hóa cao hơn và có xu hướng tập trung ở các giá trị trung bình.



- Kiểm định phân phối của các biến dữ liệu số

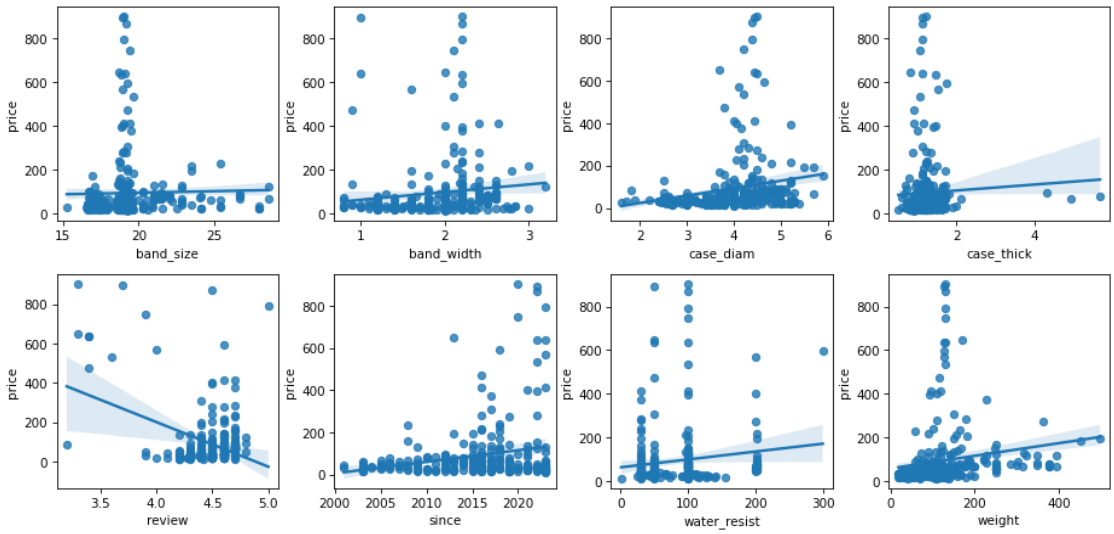
- Ta đặt ra giả thuyết H0 rằng các biến số tuân theo phân phối chuẩn. Ta sử dụng Quantile-Quantile plot để quan sát giả thuyết này.



Nhận xét:

- Dựa vào Q-Q plot, ta thấy tất cả các biến đều cách rất xa đường chuẩn (đỏ). Cho thấy các biến số không tuân theo phân phối chuẩn. Ta có thể sử dụng thêm kiểm định Anderson-Darling để có một kết luận chính xác hơn.

Sự tương quan giữa các biến số đối với biến mục tiêu 'price'



***** *Nhận xét:*

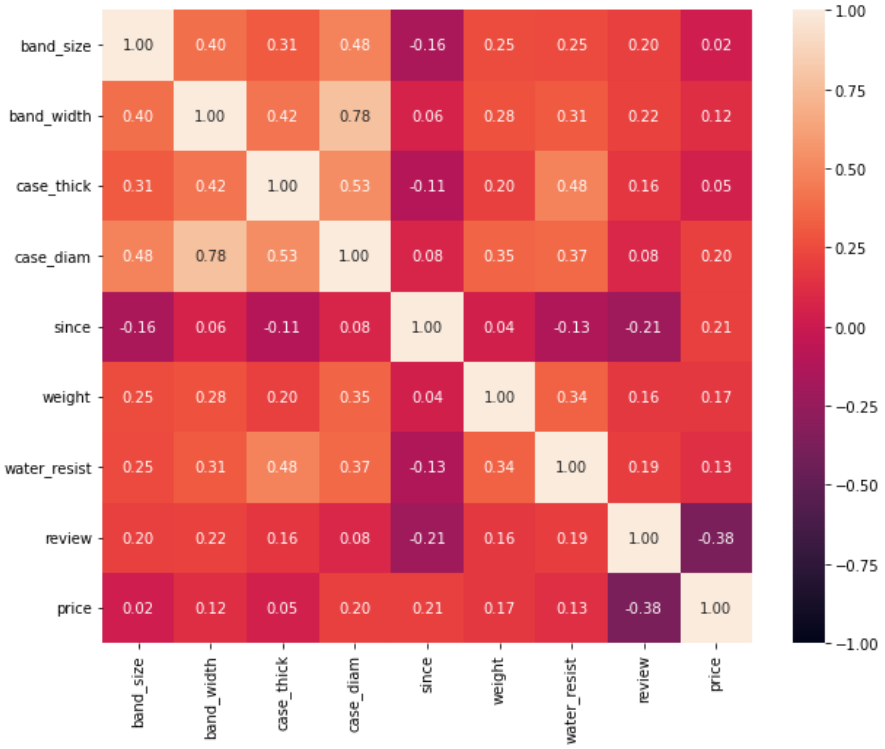
- Tất cả các biến số đều không có tương quan cao đối với biến mục tiêu 'price'. Tuy nhiên, vì số lượng biến số trong bộ dữ liệu này ít và số lượng điểm dữ liệu của từng biến số không phân bố đồng đều, chủ yếu tập trung vào các

mẫu dữ liệu có giá trị 'price' thấp nên ta không thể tùy tiện loại bỏ các biến số này được.

- **Sự tương quan giữa các biến số với nhau**

**Nhận xét:*

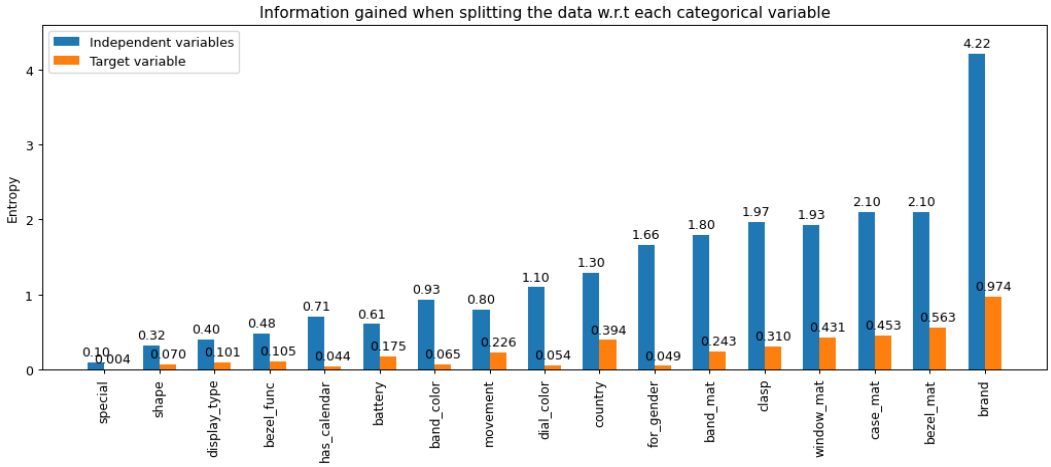
- Nhìn chung, tương quan giữa các biến số với nhau và với biến mục tiêu không quá cao. Những cặp biến số có giá trị tương quan vừa và cao bao gồm 'band_size' và 'band_width' (0.4), 'case_thick và case_diam' (0.53), 'case_diam' và 'band_width' (0.78), và các biến có tương quan cao nhất đối với biến mục tiêu là 'case_diam' (0.2) và 'since' (0.21).
- Có một số cặp biến số có giá trị tương quan âm như là: 'review' và 'price' (-0.38), 'since' và 'review' (-0.21), 'since' và 'water_resist' (-0.13). Điều này có nghĩa là các cặp biến số này có xu hướng tăng hoặc giảm ngược chiều nhau



- Có một số cặp biến số không có mối tương quan với nhau: 'price' và 'case_thick' (0.05), 'weight' và 'since' (0.04), 'review' và 'diam' (0.08)

4.2. Phân tích các biến phân loại

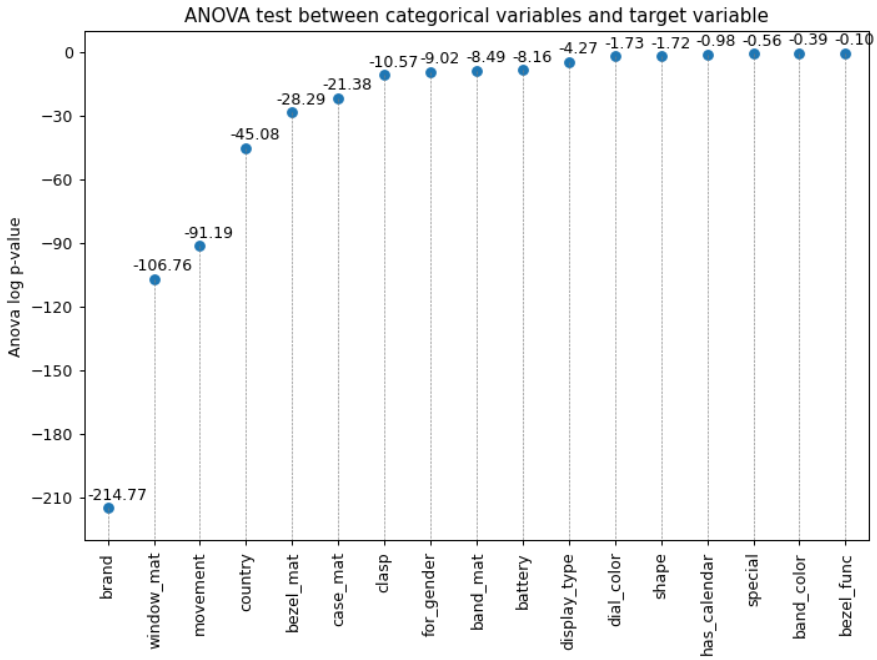
- **Lượng thông tin thu được khi phân chia dữ liệu với từng biến phân loại**



** Nhận xét:*

- Biểu đồ này thể hiện lượng thông tin (entropy) thu được khi dữ liệu được phân chia theo từng biến phân loại. Giá trị entropy càng cao thì biến đó càng có khả năng phân loại dữ liệu tốt.
- Biến 'brand' có giá trị entropy cao nhất (4.22). Điều này có nghĩa đây là biến có khả năng phân loại cao nhất. Bên cạnh đó, một số biến khác cũng có giá trị entropy tương đối cao bao gồm: 'bezel_mat' (2.1), 'case_mat'(2.1), 'window_mat'(1.93), 'clasp'(1.97), 'band_mat'(1.8)
- Một số biến có khả năng phân loại kém: 'special' (0.1), 'shape' (0.32), 'display_type' (0.4), 'bezel_func' (0.48)

- **Kết quả phân tích ANOVA giữa các biến phân loại và biến mục tiêu**



***** *Nhận xét:* (log(0.05) ≈ -3)

- Có 6 biến phân loại có p-value lớn hơn mức ý nghĩa 0.05 là 'dial_color', 'shape', 'has_calendar', 'special', 'band_color' và 'bezel_fun'. Vì 6 biến này không có ý nghĩa phân loại nên ta **lọc bỏ** các biến này.
- Các biến phân loại còn lại có giá trị p-value nhỏ hơn mức ý nghĩa 0.05. Điều này có nghĩa là với mức độ tin cậy 95%, có thể kết luận rằng các biến loại có ảnh hưởng đến mức giá của đồng hồ.
- Biến 'brand' có giá trị p-value thấp nhất (≈ 10^-93) và thấp hơn rất nhiều so với các biến phân loại khác. Vì vậy có thể nói biến 'brand' có ảnh hưởng nhất đến mức giá của đồng hồ. Theo sau là window_mat, movement, country và bezel_mat.

5. XÂY DỰNG MÔ HÌNH

5.1. Mô hình dự đoán

- Mô hình dự đoán được xây dựng dựa trên cơ sở của Linear Tree [4] bằng cách kết hợp 2 thuật toán: Linear Regression và Decision Tree. Linear Tree tương tự như Decision Tree ở chỗ nó dựa trên một tiêu chí được quy định ở nút cha để chuyển đến nút con. Khác với Decision Tree, mỗi nút lá của Linear Tree không phải là một giá trị dự đoán đơn thuần mà thay vào đó là một mô hình Linear Regression với dữ liệu huấn luyện là các điểm dữ liệu thỏa đường đi từ nút gốc đến nút lá đó.
- Mô hình được chọn để sử dụng vì những lý do sau:

a. Số lượng biến phân loại chiếm đa số: Ta cần một dạng mô hình cây quyết định để có thể tận dụng được các biến phân loại thay vì bỏ qua nó
b. Tương quan giữa các biến số thấp: các mô hình chỉ sử dụng biến số sẽ cho kết quả thấp, nên ta cần một mô hình có thể kết hợp vừa biến số và vừa biến phân loại.

- Điểm bất lợi của việc sử dụng mô hình này là yêu cầu dữ liệu cao. Mỗi nút lá cần tối thiểu 2 điểm dữ liệu để có thể dự đoán và 5-10 điểm để có ý nghĩa thống kê.
- Kết quả dự đoán của mô hình là một bộ gồm (giá trị dự đoán, khoảng tin cậy). Trong đó, khoảng tin cậy được tính với mức ý nghĩa 5%.
- Tiêu chí phân loại trên mỗi nút được chọn sao cho các mô hình ở nút con có hiệu suất hồi quy tốt nhất, tính theo R2 score.

5.2. Dự đoán với dữ liệu nhập vào không đầy đủ

- Ngoài ra, nhóm cũng hỗ trợ dự đoán khi chỉ nhập một vài giá trị của thuộc tính. Để đạt được việc đó, nhóm đã sử dụng hai công cụ là Naive Bayes và K-Nearest Neighbors.
- Ý tưởng chính của việc này là sử dụng Naive Bayes để dự đoán các giá trị biến phân loại còn thiếu, và K-Nearest Neighbors để dự đoán các biến số chưa biết.
- Kết quả của Naive Bayes là một danh sách gồm bộ thuộc tính phân loại đầy đủ và xác suất xuất hiện của bộ giá trị đó. Giá trị xác suất này sẽ có tổng bằng 1 trên toàn bộ danh sách, và được tính bằng cách coi các thuộc tính phân loại là độc lập với nhau.
- Đối với K-Nearest Neighbor, nhóm sử dụng siêu tham số K = 5, dựa trên các giá trị phân loại để tìm ra các điểm lân cận. Khoảng cách được tính là số giá trị khác nhau của giữa 2 bộ thuộc tính. Giá trị dự đoán sẽ là trung bình có trọng số các điểm tìm được, với trọng số là nghịch đảo bình phương khoảng cách.
- Việc dự đoán sẽ được thực hiện trên từng bộ thuộc tính tìm được của Naive Bayes và KNN. Kết quả trả về là bộ gồm xác suất của bộ thuộc tính tương ứng và khoảng tin cậy dự đoán được. Sau đó, các khoảng tin cậy và xác suất sẽ được kết hợp lại với nhau để cho ra khoảng tin cậy cuối cùng. Điểm chính giữa của khoảng tin cậy được chọn làm giá trị dự đoán.

6. ĐÁNH GIÁ MÔ HÌNH

6.1. Bộ dữ liệu kiểm thử và thang đo

- Bộ dữ liệu kiểm thử được trích ra từ bộ dữ liệu gốc trước khi dữ liệu được chuẩn hóa và làm sạch. Vì tính yêu cầu dữ liệu cao của mô hình, nhóm trích ra 20 điểm trong tổng số 300 điểm dữ liệu có sẵn, chiếm 6.7% bộ dữ liệu.
- Việc tách bộ dữ liệu trước khi chuẩn hóa và làm sạch là điều quan trọng, vì các giá trị biến phân loại chỉ xuất hiện 1 lần trong bộ dữ liệu huấn luyện sẽ được kết hợp lại thành một giá trị 'other' để đảm bảo yêu cầu về dữ liệu của mô hình. Nếu tách dữ luyện huấn luyện sau thì có khả năng một giá trị sẽ nằm hoàn toàn bên bộ kiểm thử, hoặc để lại 1 giá trị cho bên huấn luyện. Việc này gây ảnh hưởng lớn đến hiệu suất.
- Thang đo nhóm sử dụng là Root mean squared error (đo mức độ sai lệch giá trung bình) và R2 (đo độ hiệu quả hồi quy của mô hình).

6.2. Kết quả và nhận xét

- Thử nghiệm trên bộ dữ liệu kiểm thử cho kết quả thang đo RMSE là 33.68 và giá trị R2 là 0.702. Kết quả này cho thấy rằng mô hình dự đoán tương đối tốt các điểm dữ liệu mới.

```
test(model, df_test)

Test set cardiality: 20
Root mean squared error: 33.677; R2-score: 0.702

(33.67735982822882, 0.7017412569134351)
```

- Tuy nhiên, vì số lượng dữ liệu còn tương đối ít, nên mô hình chưa có khả năng khái quát tốt đối với các điểm dữ liệu hiếm (điểm dữ liệu mà khi đạt đến nút lá của cây thì chỉ có 2-5 điểm dữ liệu được dùng để huấn luyện). Nên kết quả dự đoán ở những điểm dữ liệu như vậy về cơ bản sẽ không được tốt như các điểm dữ liệu kia.
- Hình bên dưới cho thấy việc trích 10 điểm ngẫu nhiên trong bộ dữ liệu huấn luyện để mô hình dự đoán. Kết quả cho thấy R2 score không cao như quá trình kiểm thử.

```
test(model, df.sample(10))

Test set cardiality: 10
Root mean squared error: 43.836; R2-score: 0.218

(43.835805684394586, 0.21792141165652512)
```

TÀI LIỆU THAM KHẢO

[1] Amazon.com : watch https://www.amazon.com/s?k=watch&crid=2CBLTQTNBOYCE&sprefix=watch+%2Caps%2C745&ref=nb_sb_noss_2. Truy cập ngày 12/12/2023

[2] Sampurna Chapagain. How to Create Your Own Google Chrome Extension. Năm 2022

[3] Jason Brownlee. KNN imputation for missing values in machine learning, MachineLearningMastery.com. Năm 2020.

[4] Marco Cerliani. Linear Tree: the perfect mix of Linear Model and Decision Tree. Towards Data Science. Năm 2021

PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

STT	Thành viên	Nhiệm vụ
1	Võ Viết Thuận	- Thiết kế công cụ crawl trang Amazon - Phân tích thăm dò các biến phân loại - Thiết kế mô hình dự đoán
2	Dương Tấn Hoàng	- Phân tích thăm dò các biến số - Trình bày slide thuyết trình
3	Tăng Minh Hiển	- Crawl dữ liệu - Trình bày báo cáo
4	Phan Quốc Vỹ	- Crawl dữ liệu - Trình bày báo cáo