BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC BÀ RỊA - VŨNG TÀU**

**KHOA KỸ THUẬT - CÔNG NGHỆ**

**BÁO CÁO MÔN HỌC  
ỨNG DỤNG TRÍ TUỆ NHÂN TẠO**

**Tên đề tài:**

**SỬ DỤNG THUẬT TOÁN ABC PHÂN TÍCH DỮ LIỆU**

**GVHD:** *TS. Bùi Thị Thu Trang*

**Sinh viên:** *Họ và tên*

**Lớp:**

**Vũng Tàu, tháng 12 năm 2022**

# MỤC LỤC

# I. YÊU CẦU ĐỀ BÀI

Câu 1. Trình bày về máy học và các ứng dụng của máy học.

Câu 2. Trình bày thuật toán phân lớp và ứng dụng để phân tích dữ liệu.

# PHẦN I. MÁY HỌC VÀ ỨNG DỤNG

Trình bày tối đa 3 trang

Chú ý ko được có từ “bạn, mình, tớ,..”

Tham khảo ở đâu thì dán link vào mục Tài liệu tham khảo và đánh số thứ tự tài liệu tham khảo ở đây

# PHẦN 2. THUẬT TOÁN PHÂN LOẠI VÀ PHÂN TÍCH DỮ LIỆU

2.1. Thuật toán phân loại

KNN

LogRegression

2.2. Dữ liệu

Link dữ liệu: <https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation>

Mô tả dữ liệu: 23 thuộc tính về điểm đến, các điều kiện,…có ô dữ liệu trống (cụ thể từng thuộc tính là gì?)

Đầu ra: đề xuất cho người lái xe có giảm giá hay không.

Mô tả xem yêu cầu là gì????

Chụp hình vài dòng hình ảnh file dữ liệu ra đây!

2.2.1. Bài toán kinh doanh

Dự đoán liệu khách hàng có chấp thuận phiếu giảm giá hay không là một bài toán khó và chúng ta không thể chỉ giới thiệu nó cho tất cả mọi người vì các chi phí liên quan. Vì vậy, trong bài toán này, chúng ta sẽ dự đoán xem khách hàng sẽ chấp thuận hay từ chối phiếu giảm giá được cung cấp dựa trên hồ sợ và lịch sử khách hàng.

2.2.2 Yêu cầu

Mục tiêu của bài toán là dự đoán xem khách hàng sẽ chấp nhận hay từ chối phiếu giảm giá một địa điểm cụ thể dựa trên các thuộc tính. Nếu khách hàng chấp nhận phiếu giảm giá sẽ dán nhãn Y=1 và nếu khách hàng từ chối phiếu giảm giá sẽ được đánh giá Y=0.

2.2.3 Dữ liệu

Dữ liệu này được thu thập thông qua một cuộc khảo sát trên Amazon Mechanical Turk được xuất bản trên trang web UCI vào năm 2017. Cuộc khảo sát mô tả các tình huống lái xe khác nhau bao gồm điểm đến người dùng, thời gian, thời tiết, hành khách, thuộc tính phiếu giảm giá,… Sau đó hỏi xem người dùng có chấp thuận phiếu giảm giá không.

Tệp dữ liệu: in-vehicle-coupon-recommendation.csv

Bộ dữ liệu bao gồm hơn 12684 điểm dữ liệu và 26 thuộc tính

Giải thích về các thuộc tính/ tính năng của bộ dữ liệu:

Thuộc tính người dùng:

* Gender: Female, Male
* Age:21, 46, 26, 31, 41, 50plus, 36, below21
* Marital Status: Unmarried partner, Single, Married partner, Divorced, Widowed
* has\_Children: 1: has children, 0: No children
* Education: Some college — no degree, Bachelors degree, Associates degree, High School Graduate, Graduate degree (Masters or Doctorate), Some High School
* Occupation: nghề nghiệp của người dùng (Unemployed, Architecture & Engineering, Student,Education&Training&Library, Healthcare Support,Healthcare Practitioners & Technical, Sales & Related, Management,Arts Design Entertainment Sports & Media, Computer & Mathematical,Life Physical Social Science, Personal Care & Service, Community & Social Services, Office & Administrative Support, Construction & Extraction, Legal, Retired, Installation Maintenance & Repair, Transportation & Material Moving, Business & Financial, Protective Service, Food Preparation & Serving Related, Production Occupations, Building & Grounds Cleaning & Maintenance, Farming Fishing & Forestry)
* Income: thu nhập của người dùng (Less than $12500,$12500 — $24999,$25000 — $37499,$37500 — $49999,$50000 — $62499,$62500 — $74999,$75000 — $87499,$87500 — $99999,$100000 or More)
* Car : Mô tả phương tiện do người dùng điều kiển (Scooter and motorcycle, crossover, Mazda5) (99% of values are missing in this feature)
* Bar: người dùng đến bar bao nhiêu lẩn mỗi tháng (never, less1, 1~3, 4~8, gt8, nan)
* CoffeeHouse: người dùng đến quán cà phê bao nhiêu lần mỗi tháng (never, less1, 1~3, 4~8, gt8, nan)
* CarryAway: người dùng nhận đồ ăn mang về bao nhiêu lần mỗi tháng (never, less1, 1~3, 4~8, gt8, nan)
* RestaurantLessThan20: bao nhiêu lần người dùng đến nhà hàng với chi phí trung bình cho mỗi người dưới 20$ mỗi tháng? (never, less1, 1~3, 4~8, gt8, nan)
* Restaurant20To50: người dùng đến nhà hàng bao nhiêu lần với chi phí trung bình cho mỗi ngườ là 20 – 50$ mỗi tháng? (never, less1, 1~3, 4~8, gt8, nan)

Thuộc tính ngữ cảnh:

* Destination: điểm đến của người dùng (No Urgent Place, Home, Work)
* Passenger: khách hàng trên xe là ai (Alone, Friend(s), Kid(s), Partner)
* Weather: thời tiết khi người dùng đang lái xe (Sunny, Rainy, Snowy)
* Temperature: nhiệt độ tính bằng độ F khi người dùng lái xe (55, 80, 30)
* Time: thời gian khi người dùng lái xe (2PM, 10AM, 6PM, 7AM, 10PM)
* toCoupon\_GEQ5min: khoảng cách lái xe đến nhà hàng/quán cà phê/ quán bar để sử dụng phiếu giảm giá lớn hơn 5 phút (0,1)
* toCoupon\_GEQ15min: khoảng cách lái xe đến nhà hàng/quán cà phê/ quán bar để sử dụng phiếu giảm giá lớn hơn 15 phút (0,1)
* toCoupon\_GEQ25min: khoảng cách lái xe đến nhà hàng/quán cà phê/ quán bar để sử dụng phiếu giảm giá lớn hơn 25 phút (0,1)
* direction\_same: liệu nhà hàng/ quán cà phê/ quán bar có cùng hướng với điểm đến hiện tại của người dùng hay không (0,1)
* direction\_opp: liệu nhà hàng/ quán cà phê/ quán bar có ở hướng ngược lại với điểm đến hiện tại của người dùng hay không (0,1)

Thuộc tính phiếu giảm giá:

* Coupon: loại phiếu giảm giá do công ty cung cấp (Restaurant(<$20), Coffee House, Carry out & Take away, Bar, Restaurant($20-$50))

Ở đây, Restaurant(<$20) là mức trả trung bình cho mỗi người dùng dưới $20, nhà hàng không quá đắt. Restaurant($20-$50) có nghĩa là mức trả trung bình cho mỗi người dùng là từ $20 đến $50, nhà hàng hơi đắt tiền.

* Expiration: Phiếu giảm giá hết hạn sau 1 ngày hoặc sau 2h (1d, 2h)

Thuộc tính đích:

* Y: phiếu giảm giá được chấp thuận hay từ chối, 1: được chấp thuận , 0: bị từ chối

2.3. Phân tích dữ liệu

Import thư viện gì? Dùng để làm gì?

Xử lý các thuộc tính như thế nào?

Dùng thuật toán gì để xây dựng model?

Độ chính xác của model trên tập train và test

Nếu dùng nhiều thuật toán khác nhau thì so sánh độ chính xác và thời gian trình bày trong bảng biểu

A screenshot of a computer

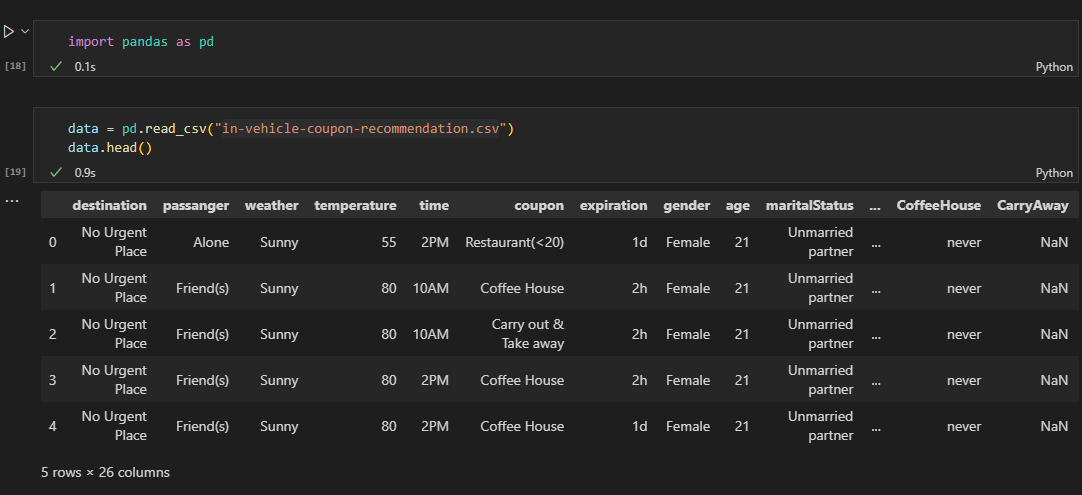
Description automatically generated with medium confidence

- Thư viện pandas để đọc, ghi, xử lý file dữ liệu

- Thư viện matplotlib dùng để vẽ đồ thị

2.3.1 Tải dữ liệu

Để tải dữ liệu, chúng ta chỉ cần tệp in-vehicle-coupon-recommendation.csv và tải tệp này bằng khung dữ liệu pandas.



Text

Description automatically generated

Bộ dữ liệu này có 12684 dòng dữ liệu và 26 cột cùng với nhãn mục tiêu của chúng. 5 hàng đầu tiên xuất hiện như hình trên.

2.3.2 Phân bổ coupon

Text

Description automatically generated

Tập dữ liệu này được cân bằng một phần với nhãn loại chấp nhận là khoảng 57% và loại từ chối là khoảng 43%

2.3.3 Loại bỏ trùng lặp

Text

Description automatically generated

2.3.4 Kiểm tra các giá trị còn thiếu trong tệp dữ liệu

Text

Description automatically generated

Chúng ta có thể thấy rằng có tổng 6 thuộc tính giá trị bị thiếu, từ tính năng ‘car’ có 99% giá trị bị thiếu nên chúng ta cần bỏ đặc trưng này vì ngay cả sau khi dự đoán các giá trị bị thiếu, đặc trưng này ít quan trọng hơn và nó ít có khả năng dự đoán hơn.

Các đặc trưng khác như ‘Bar’, ‘CoffeeHouse’, ‘CarryAway’, ‘RestaurantLessThan20’, ‘Restaurant20To50’ có khoảng 1% giá trị bị thiếu, vì vậy chúng tôi cần điền các giá trị bị thiếu này bằng 1 số giá trị khác, vì vậy chúng tôi sẽ sử dụng mode imputation.

2.3.5 Điền các giá trị bị thiếu bằng cách sử dụng mode imputation

Mode imputation thay thế các giá trị còn thiếu của một đối tượng phân loại bằng giá trị phổ biến nhất của đối tượng đó.

Text

Description automatically generated

2.3.6 Mối tương quan giữa các đặc trưng

A picture containing graphical user interface

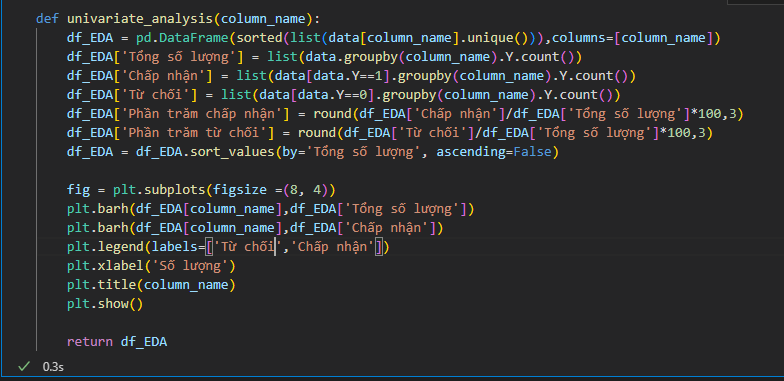
Description automatically generated

Từ ma trận tương quan, chúng ta có thể thấy rằng đặc trưng ‘direction\_same’ tương quan hoàn hảo với ‘direction\_opp’, cả hai đều có cùng giá trị tương quan với thuộc tính đích. Vì vậy không cần 2 tính năng tương quan hoàn hảo để chúng tôi có thể loại bỏ một trong số chúng.

Đặc trưng ‘toCoupon\_GEQ5min’ không tương quan với thuộc tính đích vì đặc trưng này có cùng giá trị ‘1’ cho tất cả dòng dữ liệu, có nghĩa tất cả các điểm ăn uống đều các người dùng ít nhất 5 phút.

Vì vậy, cần bỏ cả hai đặc tính ‘direction\_opp’ và ‘toCoupon\_GEQ5min’ vì đạc tính này ít quan trọng hơn và sẽ không giúp ích nhiều cho việc dự đoán.

2.3.7 Phân tích đơn biến

Vì tất cả các thuộc tính đều có tính phân loại, nên chúng ta sẽ xem xét các phân phối kín đáo của các biến khác nhau với biến mục tiêu. 

1. Coupon:

Text

Description automatically generated

Chart, bar chart

Description automatically generated

A picture containing text, monitor, several

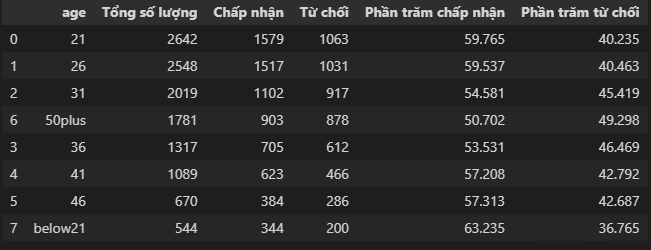
Description automatically generated

* Chúng ta có thể thấy rằng các phiếu giảm giá được cung cấp nhiều nhất là dành cho Coffee House. Ngoài ra, phiếu giảm giá được nhiều người chấp nhận là phiếu giảm giá ‘Carry out & Take away’ với 73% và ‘Restaurant(<20)’ là 70%. Phiếu giảm giá có thỉ lệ chấp thuận thấp nhất là ‘Bar’

1. Age:

Chart, bar chart

Description automatically generated



* Chúng ta có thể thấy rằng hầu hết người dùng trong dữ liệu này có độ tuổi từ 21 đến 31 tuổi. Ngoài ra, người dùng có độ tuổi dưới 21 có tỷ lệ chấp nhận phiếu giảm giá cao nhất. Ngoài ra, người dùng trên 50 tuổi có tỷ lệ từ chối phiếu giảm giá cao nhất

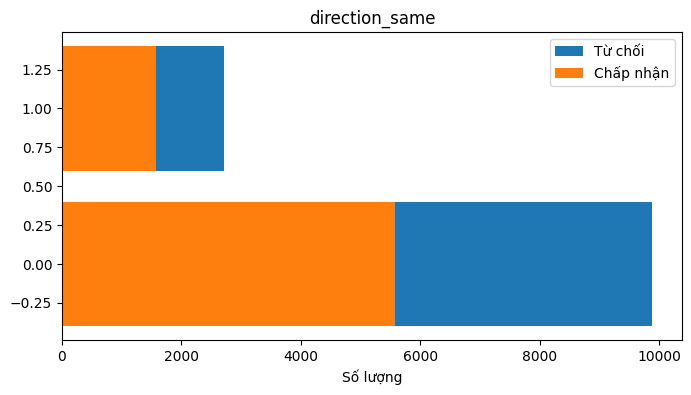
1. Income: Chart, bar chart

   Description automatically generated

A black screen with white text

Description automatically generated with low confidence

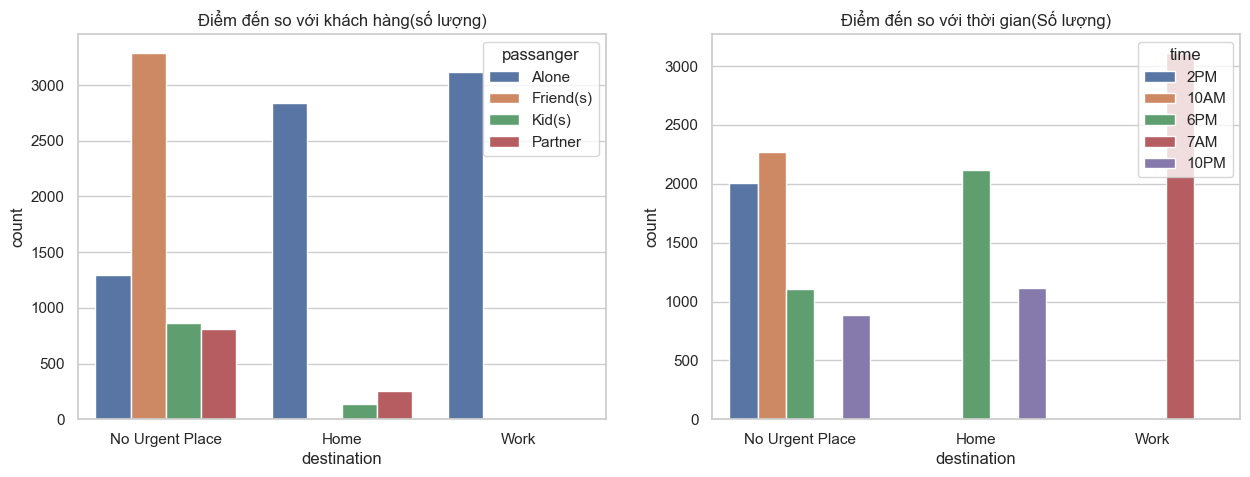
* Hầu hết người dùng trong dữ liệu này có thu nhập từ 12500 đô đến 49999 đô. Người dùng có thu nhập cao chấp nhận nhiều phiếu giảm giá hơn những người khác. Ngoài ra người có thu nhập tầm trung hầu hết từ chối phiếu giảm giá.

1. direction\_same: Graphical user interface, application, website

   Description automatically generated

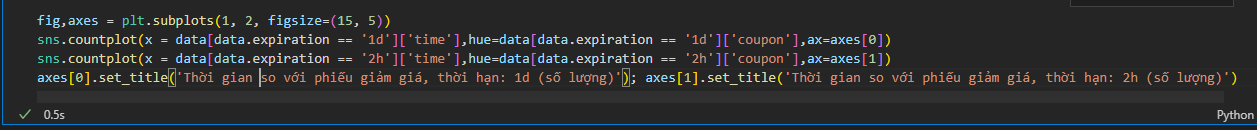
* Đặc tính ‘direction\_same’ có hai giá trị là 0 và 1. Cả hai giá trị này có tỷ lệ chấp nhận và từ chối gần như tương tự nhau. Vì vậy đặc tính này không hữu ích trong dự đoán. Vì vậy, bỏ đặc tính này.
  + 1. Phân tích hai biến số

1. Người dùng đi đến ‘Work’, ‘Home’, ‘No Urgent Place’ lúc nào , với ai ?Text

   Description automatically generatedA black screen with white text

   Description automatically generated with low confidence

* Người dùng đi cùng bạn bè ‘Friend(s)’ chỉ có điểm đến không khẩn cấp ‘No Urgent Place’.
* Vào lúc 7AM, chỉ người dùng đi làm ‘Work’ mới ra ngoài một mình.
* Vào lúc 2PM, người dùng chỉ đến nơi không khẩn cấp ‘No Urgent Place’
* ‘Home’, người dùng chủ yếu chỉ đi 1 mình vào lúc 6PM và 10PM
* ‘No Urgent Place’, người dùng chủ yếu đi một mình ‘Alone’ lúc 2PM và với bạn ‘Friend(s)’ lúc 6PM
* Người dùng đi cùng bạn bè ‘Friend(s)’ lúc 6PM có tỉ lệ chấp nhận cao nhất
* Người dùng đi một mình ‘Alone’ lúc 2PM cũng có tỉ lệ chấp nhận cao
* Người dùng đi cùng trẻ em lúc 10PM hầu hết từ chối phiếu giảm giá

1. Tỷ lệ chấp nhận và từ chối phiếu giảm giá cao vào thời điểm nàoChart, bar chart

   Description automatically generatedA black screen with white text

   Description automatically generated with low confidence

* Vào lúc 7AM và 2PM, người dùng sẽ không nhận phiếu giảm giá bar có thời hạn 2h
* Vào lúc 2PM và 6PM, người dùng chủ yếu chấp nhận phiếu giảm giá ‘Restaurant(<20)’ và ‘Carry out & take away’ với tỉ lệ 80%
* Vào lúc 2PM, người dùng chủ yếu chấp nhận phiếu giảm giá ‘Carry out & take away’ với tỉ lệ 86%
* Từ 7AM đến 6PM. Phiếu giảm giá Coffee Hosue là nhiều nhất
* 10PM chỉ có Carry out & take away(1d) và Restaurant(<20)(2h) là cao nhất.
* Vào 10PM, người dùng không chấp nhận phiếu giảm giá ‘Restaurant(20-50)’
* Vào 7AM, 10AM, 2PM người dùng từ chối phiếu giảm giá cho Bar
  + 1. Thiết kế đặc trưng

1. to\_coupon

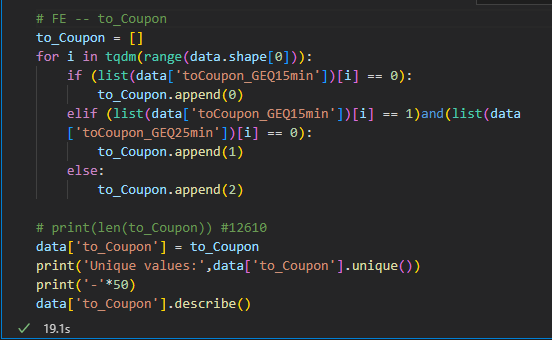
‘toCoupon\_GEQ15min’ và ‘toCoupon\_GEQ25min’ cho biết khoảng cách lái xe đến nhà hàng/quán cà phê/quán bar, vì vậy chúng tôi có thể kết hợp chúng và tạo một đặc trưng mới gọi là 'to\_coupon'.

Các đặc trưng 'toCoupon\_GEQ15min' và 'toCoupon\_GEQ25min' có hai giá trị phân loại 0 & 1 và 'to\_coupon' có ba giá trị phân loại 0,1 & 2.

0: khoảng cách lái xe nhỏ hơn hoặc bằng 15 phút

1: khoảng cách lái xe lớn hơn 15 phút và nhỏ hơn hoặc bằng 25 phút

2: khoảng cách lái xe lớn hơn 25 phút

 Text

Description automatically generated

1. coupon\_freq

coupon\_freq' là sự kết hợp của năm đặc trưng, RestaurantLessThan20, CoffeeHouse, CarryAway, Bar và Restaurant20To50. Năm đặc trưng này cho biết tần suất người dùng đến RestaurantLessThan20, CoffeeHouse, CarryAway, Bar và Restaurant20To50.

Mỗi người dùng được yêu cầu một phiếu giảm giá, vì vậy chúng tôi chỉ cần tần suất của một loại phiếu giảm giá đó, không cần các chi tiết phiếu giảm giá khác. vì vậy chúng tôi tạo một tính năng mới cho biết tần suất người dùng sử dụng phiếu giảm giá đã hỏi.

* + 1. Mã hóa dữ liệu

Hầu hết các thuật toán học máy không thể hoạt động trực tiếp trên dữ liệu phân loại. Chúng yêu cầu tất cả các biến đầu vào và biến đầu ra phải là số. Hầu hết các hiệu suất của thuật toán khác nhau dựa trên phương pháp mã hóa nào được sử dụng. Có nhiều cách để mã hóa dữ liệu phân loại.

1. Mã hóa tần số (Frequency Encoding)

Giá trị thay thế nó là tần suất giá trị đó xuất hiện ttong cột

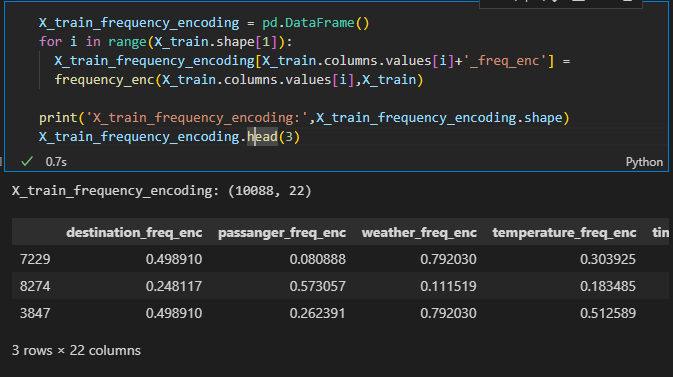
Trong mã hóa này, chúng tôi sẽ chuyển đổi các đặc trưng phân loại thành giá trị tần số của các danh mục của chúng.

Dưới đây là các bước để tính giá trị tần số của tính năng:

1. Nhóm theo biến phân loại và thu được số lượng của từng loại

2. Để chuẩn hóa chúng chia với tổng số của một biến phân loại

3. Bản đồ giá trị tần số với các tính năng đào tạo

 A screenshot of a computer

Description automatically generated with medium confidence

* + 1. Models

Sau khi thực hiện phân tích dữ liệu thăm dò và làm sạch dữ liệu, chúng tôi cũng đã thực hiện một số kỹ thuật đặc trưng và thêm đặc trưng mới 'to\_coupon', 'coupon\_freq' và 'occupation\_class' từ các đặc trưng hiện có. Chúng tôi cũng đã chuyển đổi các tính năng phân loại thành các tính năng số bằng cách sử dụng năm phương pháp mã hóa khác nhau và chuẩn bị năm ma trận dữ liệu kiểm tra và huấn luyện khác nhau. Hiện chúng tôi đã sẵn sàng áp dụng thuật toán học máy cho ma trận dữ liệu đã chuẩn bị của mình.

Có nhiều thuật toán học máy để thử nghiệm khi bạn đang giải quyết vấn đề phân loại. Bạn không bao giờ biết trước mô hình phân loại nào sẽ phù hợp nhất với mình. Vì vậy, chúng tôi sẽ thử nghiệm với một vài mô hình phân loại và kiểm tra xem mô hình nào hoạt động tốt nhất. Việc thực hiện điều chỉnh siêu tham số để tối đa hóa hiệu suất của mô hình cũng rất quan trọng.

1.Logistic Regression

2.K-Nearest Neighbor

3.Decision Tree

Ở đây chúng ta sẽ tạo một hàm model\_training. Hàm đó nhận hai đối số tên mô hình và loại mã hóa, đồng thời thực hiện đào tạo mô hình với điều chỉnh siêu tham số với ma trận dữ liệu được mã hóa và trả về siêu tham số tốt nhất, đồng thời tính toán điểm logloss và auc huấn luyện và kiểm tra.

# TÀI LIỆU THAM KHẢO