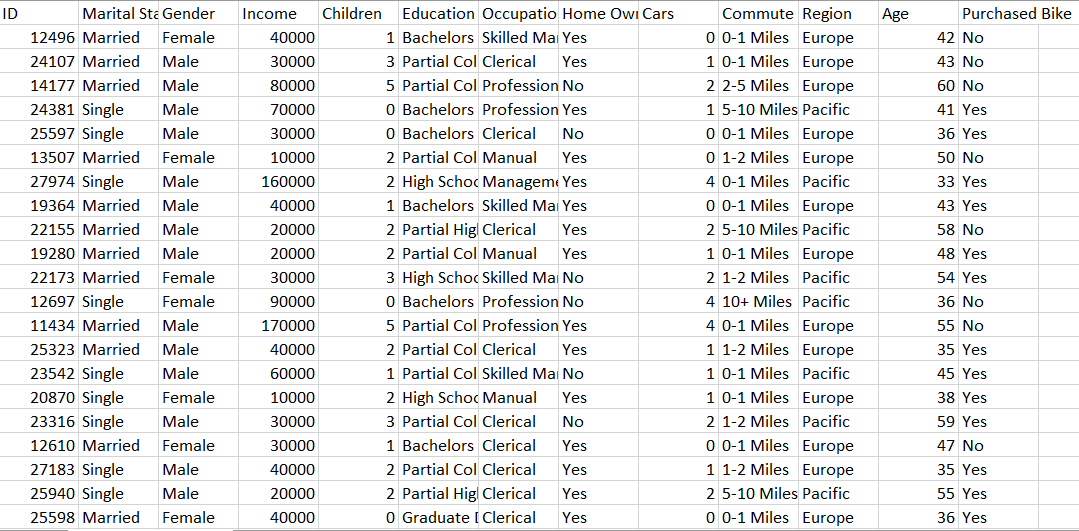
1.1. Mô tả bộ dữ liệu

- Bộ dữ liệu bike\_buyers\_clean.csv gồm 13 thuộc tính (13 cột) và 1001 dòng dữ liệu:



- Link tải:

<https://www.kaggle.com/heeraldedhia/bike-buyers>

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Mô tả** |
| 1 | ID | ID của khách hàng |
| 2 | Marital Status | Tình trạng hôn nhân |
| 3 | Gender | Giới tính |
| 4 | Income | Thu nhập |
| 5 | Children | Số con |
| 6 | Education | Tình trạng học vấn |
| 7 | Occupation | Nghề nghiệp |
| 8 | Home Owner | Nhà đã có |
| 9 | Cars | Số ô tô đã có |
| 10 | Commute Distance | Khoảng cách đến công ty |
| 11 | Region | Khu vực sinh sống |
| 12 | Age | Tuổi tác |
| 13 | Purchased Bike | Mua xe |

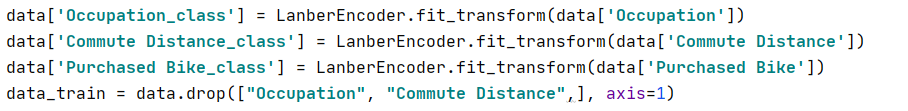
1.2. Xác định bài toán học máy

- Với bộ dữ liệu này, ta sử dụng bài toán phân lớp nhiều lớp (2 lớp)

* Đầu vào X là 12 thuộc tính (cột) đầu tiên: ID, Marital Status, Gender, Income, Children, Education, Occupation, Home Owner, Cars, Commute, Region, Age,
* Đầu ra y là thuộc tính (cột) cuối cùng: Purchased Bike
* Áp dụng bài toán phân lớp sẽ y chia thành 2 lớp: Yes, No

- Xử lý dữ liệu: Vì dữ liệu chủ yếu là dạng text, ta sử dụng các kỹ thuật tiền sử lý dữ liệu:

* LabelEncoder để mã hóa về số để tính toán



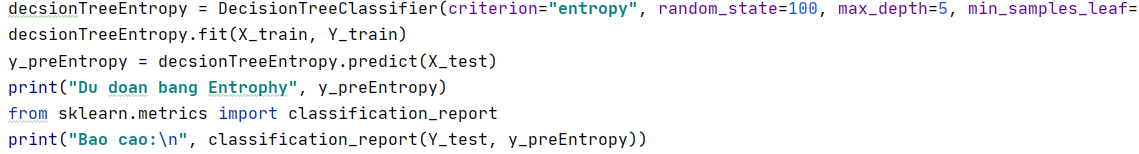
* Chia tập train và test:



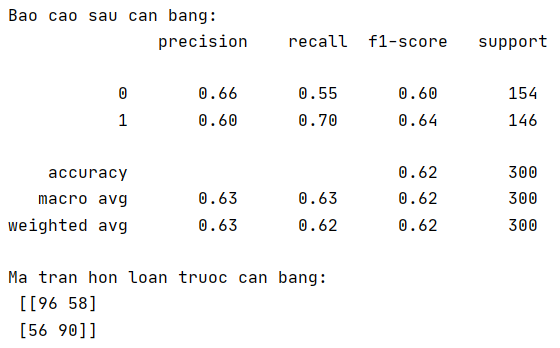
- Các kỹ thuật học máy được sử dụng: Decision Tree Classifier, Entropy

1.3. Phương pháp đáng giá của các kỹ thuật máy

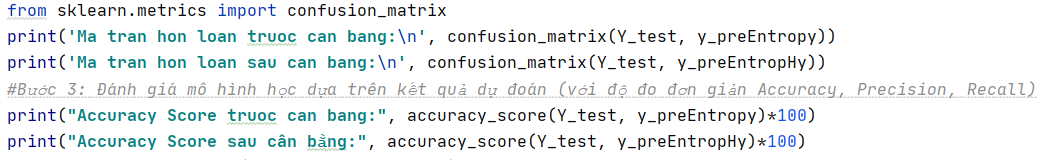
- Kỹ thuật học máy Decision Tree Classifier sử dụng phương pháp đáng giá: accuracy\_score, confusion\_matrix, classification\_report



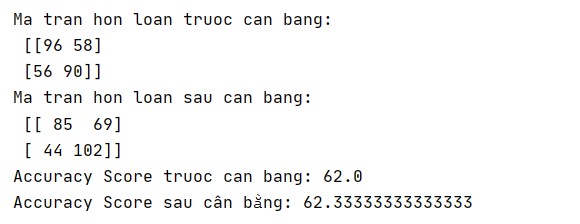
Kết quả chạy:



* Ma trận hỗn loạn sử dụng phương pháp đánh giá



Kết quả chạy



1.4. Kết luận

- Kỹ thuật phân lớp Random Forest Classifier nhạy cảm với dữ liệu không cân bằng