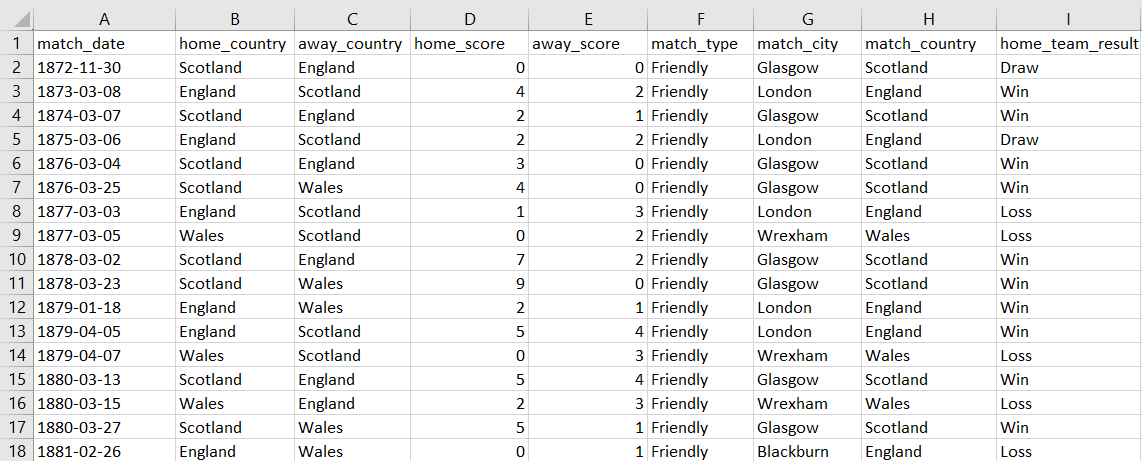
1.1. Mô tả bộ dữ liệu

- Bộ dữ liệu soccer\_international\_history.csv gồm 9 thuộc tính (9 cột) và 38686 dòng dữ liệu:



- Link tải: https://www.mldata.io/dataset-details/soccer\_international\_history/?fbclid=IwAR0vpWNgJktDZbAnsU2Gm3TWLmEaQ90Hxt5wODF1pxTsxR3oSzAxGo5Pfmc

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Mô tả** |
| 1 | match\_date | Ngày, tháng, năm thi đấu |
| 2 | home\_country | Tên nước chủ nhà |
| 3 | away\_country | Tên nước đội khách |
| 4 | home\_score | Sô bàn thắng của đội nhà |
| 5 | away\_score | Số bàn thắng đội khách |
| 6 | match\_type | Loại thi đấu: Friendly (Giao hữu) |
| 7 | match\_city | Tên thành phố (nơi thi đấu) |
| 8 | match\_country | Tên nước để đối chiếu với kết quả thi đấu |
| 9 | home\_team\_result | Kết quả thi đấu: Win (thắng), Draw (hòa), Loss (thua) |

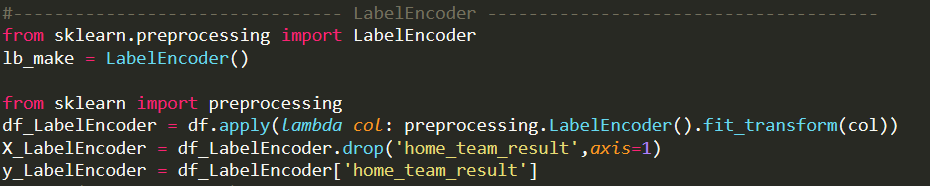
1.2. Xác định bài toán học máy

- Với bộ dữ liệu này, ta sử dụng bài toán phân lớp nhiều lớp (3 lớp)

* Đầu vào X là 8 thuộc tính (cột) đầu tiên: match\_date, home\_country, away\_country, home\_score, away\_score, match\_type, match\_city, match\_country
* Đầu ra y là thuộc tính (cột) cuối cùng: home\_team\_result
* Áp dụng bài toán phân lớp sẽ y chia thành 3 lớp: Win, Draw, Loss

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

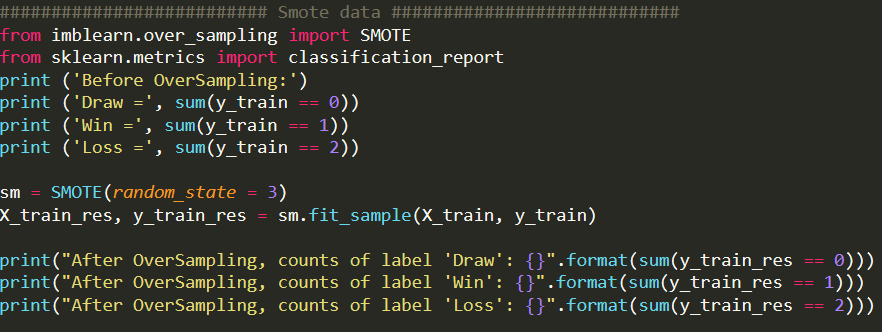
* LabelEncoder để mã hóa về số để tính toán



* Chia tập train và test:



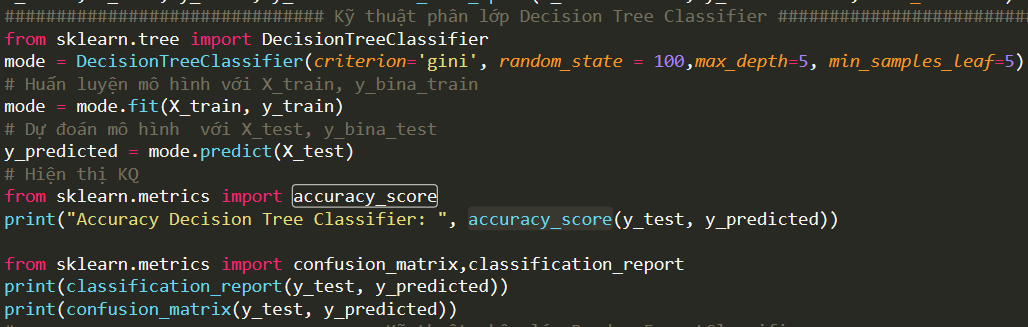
* SMOTE để cân bằng dữ liệu:



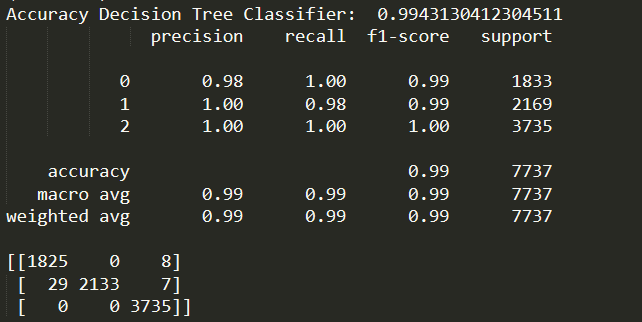
- Các kỹ thuật học máy được sử dụng: Decision Tree Classifier, Random Forest Classifier

1.3. Phương pháp đáng giá của các kỹ thuật máy

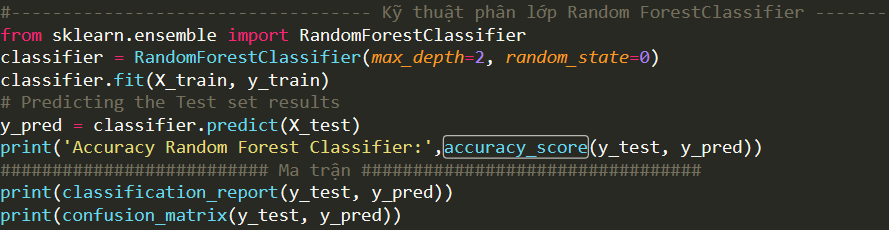
- Kỹ thuật học máy Decision Tree Classifier sử dụng phương pháp đáng giá: accuracy\_score, confusion\_matrix, classification\_report



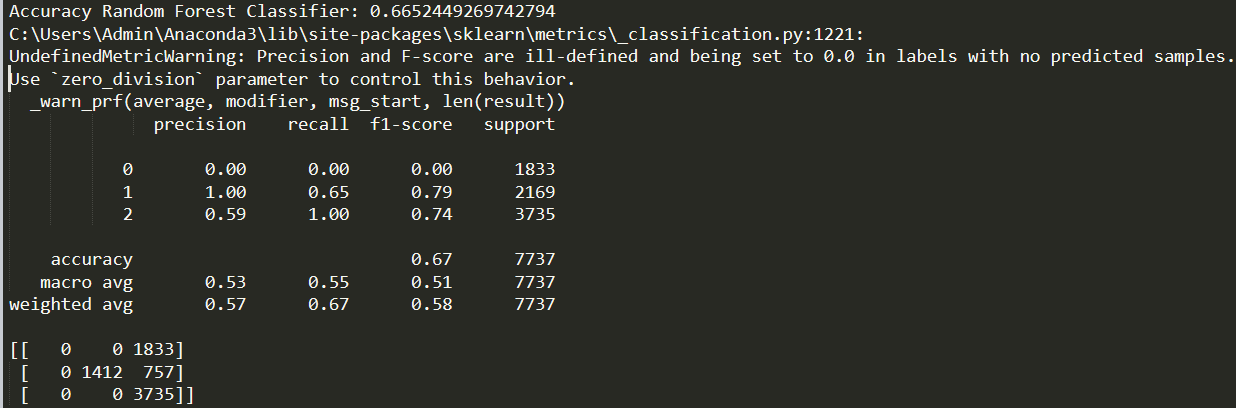
Kết quả chạy:



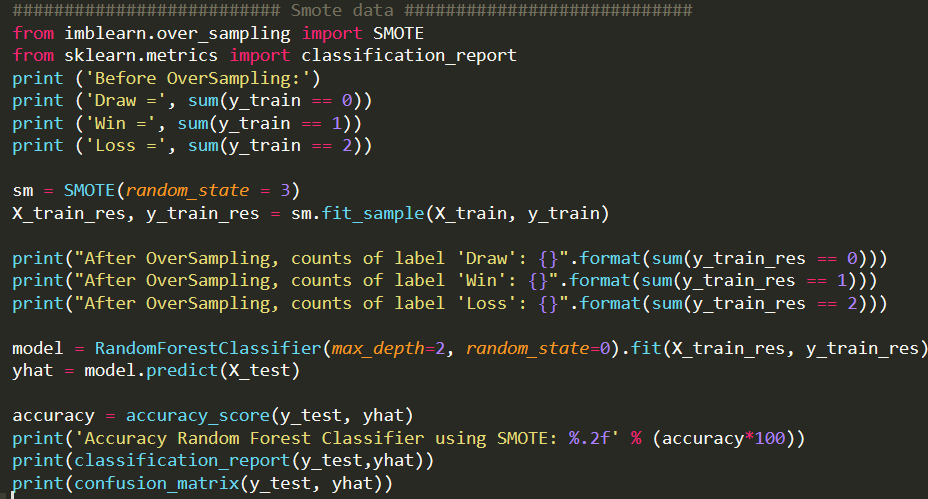
* Kỹ thuật Random Forest Classifier sử dụng phương pháp đánh giá:



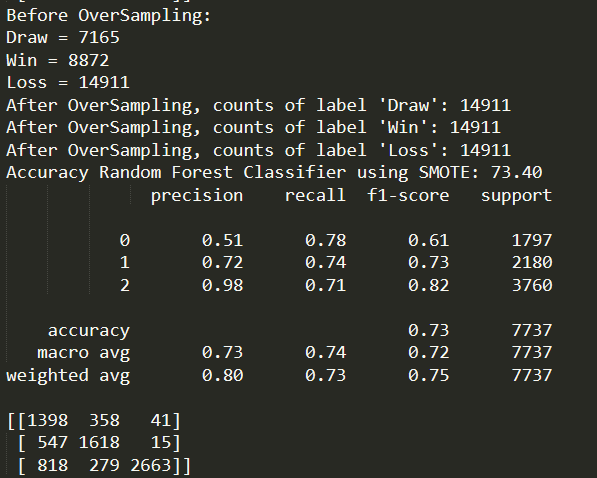
Kết quả chạy



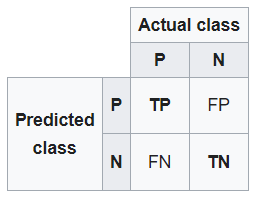
- Độc chính xác của kỹ thuật Random Forest Classifier tăng lên sau khi SMOTE (cân bằng dữ liệu):



Kết quả chạy:



* Phương pháp đáng giá confusion\_matrix và một số công thức tính:



Trong đó:

P = Positive;

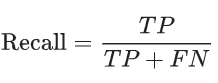
N = Negative;

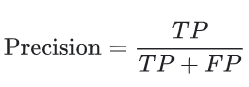
TP = True Positive;

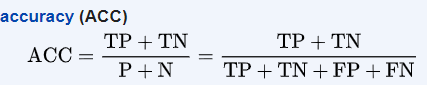
FP = False Positive;

TN = True Negative;

FN = False Negative.







1.4. Kết luận

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