**MÔ TẢ DỰ ÁN DỰ ĐOÁN**

**ĐIỂM THI TỐT NGHIỆP**

**BẰNG**

**RANDOM FOREST REGRESSOR**

* **Các file dữ liệu:**
* File dataOfficial.xlsx: dữ liệu gồm 1042 dòng, mỗi dòng là điểm tổng kết cuối năm THPT tương ứng với 9 môn Toán, Lý, Hóa, Sinh, Văn, Sử, Địa, GDCD và điểm thi tốt nghiệp tương ứng với 9 môn đó kèm theo hạnh kiểm, học lực và có đủ điều kiện để thi tốt nghiệp hay không.
* File DiemchuanOfficial.xlsx: dữ liệu gồm 306 dòng tương ứng với điểm chuẩn các khối thi của các trường đại học trong khối Đại học Quốc gia Thành phố Hồ Chí Minh.
* **Các bước thực hiện:**

1. Tạo dữ liệu để train:

* Do dữ liệu để đưa vào train ban đầu là không có (do không thu thập được) nên ta phải tiến hành tạo dữ liệu train, cách tạo như sau:
* Đối với mỗi môn, dựa vào phân phối phổ điểm của kỳ thi THPT 2018 của từng môn trong 9 môn trên, ta tạo điểm thi tốt nghiệp mỗi môn dựa vào phân phối trong phổ điểm, trong quá trình tạo có sử dụng các hàm mround(a,b) trong excel để làm tròn số a (ở đây a chính là điểm tổng kết của một môn học) đến hàng b và các phép thêm bớt thích hợp để tạo ra điểm c chính là điểm dùng để train.

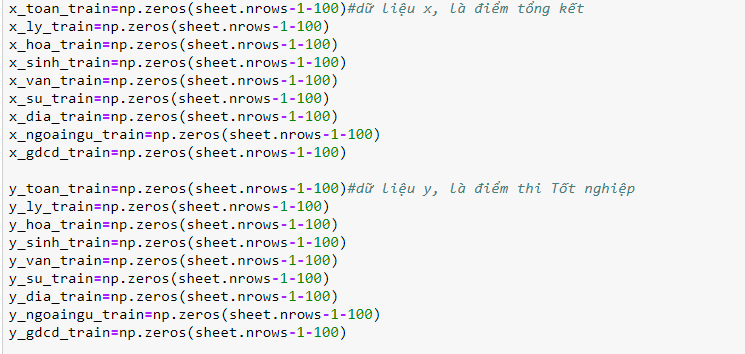
1. Đưa dữ liệu vào và train (số dữ liệu train: 942, số dữ liệu test:100):

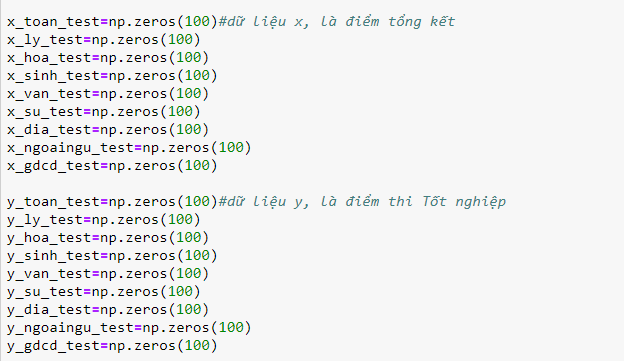
* Sử dụng thư viện xlrd để đọc file excel và thư viện numpy để thao tác với các mảng số.
* Sau đó ta đọc dữ liệu bằng 3 dòng sau:



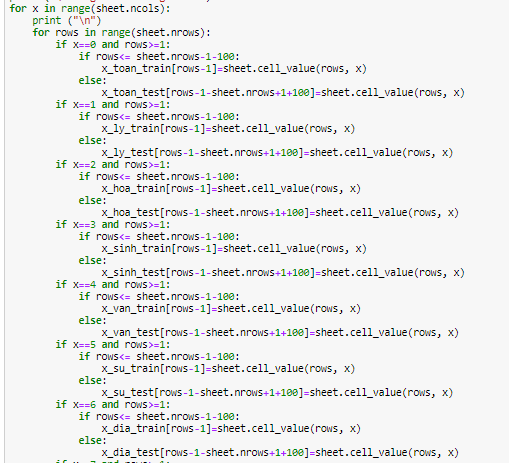
Lúc này sheet chính là khối dữ liệu của ta.

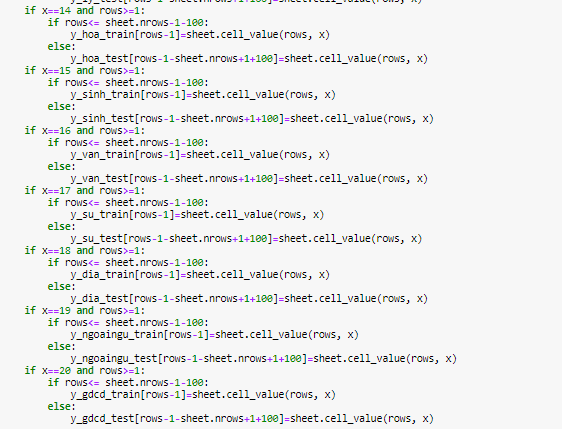
* Tiếp theo, ta tiến hành khởi tạo các mảng gồm các số 0 để chứa dữ liệu trong bảng từ sheet, ở đây X chính là dữ liệu độc lập (là điểm tổng kết năm học), còn Y chính là dữ liệu phụ thuộc (là điểm thi tốt nghiệp) lần lượt cho 9 môn học.



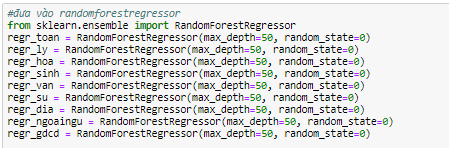


* Tiếp theo, ta lần lượt gán các giá trị trong bảng dữ liệu vào các mảng đã tạo:

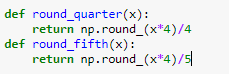




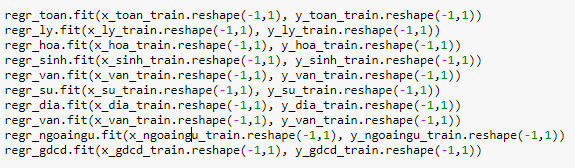
* Tiến hành import module RandomForestRegressor từ thư viện sklearn.ensemble và tạo các regressor với các thông số tương ứng cho 9 môn học.



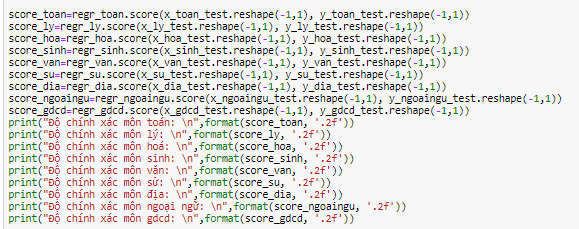
* Ở đây ta định nghĩa 2 hàm round\_quarter(x) và round\_fifth(x) để làm tròn x tương ứng đến 0.25 hoặc 0.2:



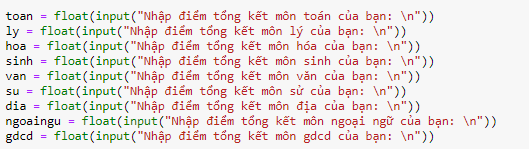
* Tiến hành fit các dữ liệu đã tạo vào các regressor, ở đây các vector ta phải reshape(-1,1) để làm phẳng chúng để tránh bị lỗi.



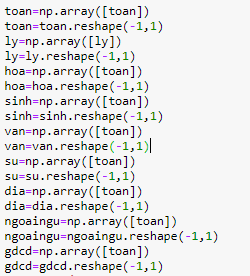
* Kiểm thử độ chính xác với mảng dữ liệu test:



1. Đọc dữ liệu của người dùng nhập vào:



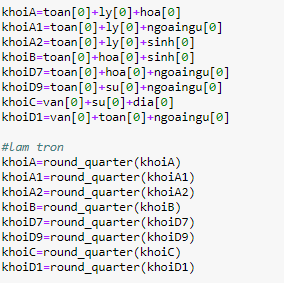
* Tiến hành đưa các dữ liệu người dùng nhập vào về mảng và reshape lại chúng:



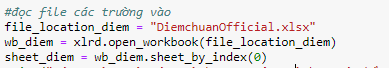
* Sử dụng hàm predict của các regressor để dự đoán từ dữ liệu người dùng nhập vào và in ra:



* Tính tổng điểm các khối thi và làm tròn tổng điểm:



* Đọc dữ liệu từ file chứa điểm chuẩn:



* Tiến hành duyệt dữ liệu file điểm chuẩn và so sánh tổng điểm dự đoán các khối thi của người dùng, nếu điểm thi của người dùng lớn hơn hoặc bằng điểm chuẩn đọc vào thì tiến hành in ngành và trường tương ứng trúng tuyển của người dùng:

