Báo cáo thực hành phân lớp với sklean (Titanic)

Họ và tên: Nguyễn Hữu Vượt

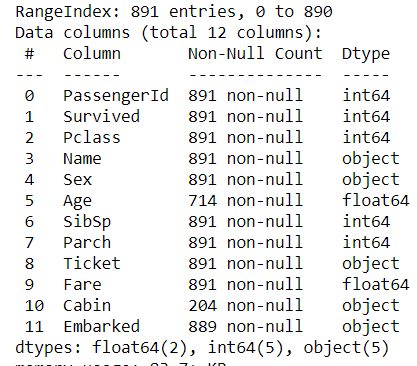
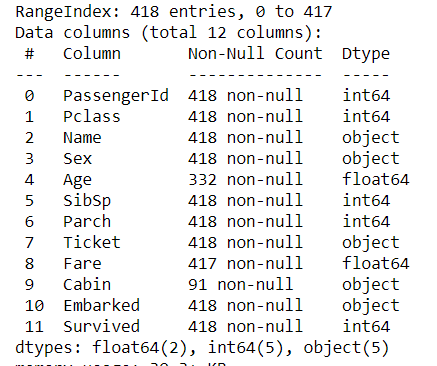
Mã sinh viên: 19020489

Nhóm : 3

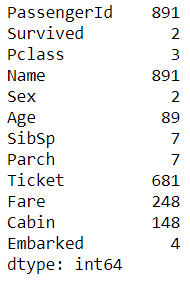
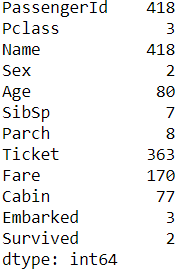
# Data Preprocessing

1. Nhận xét data ban đầu

* Train data có 12 chiều , gồm 891 dòng và 12 cột
* Test\_data có 12 chiều , gồm 418 dòng và 12 cột



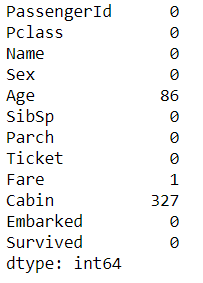
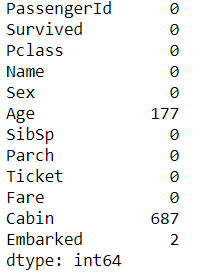
* Các đặc trưng của bộ dữ liệu train và test

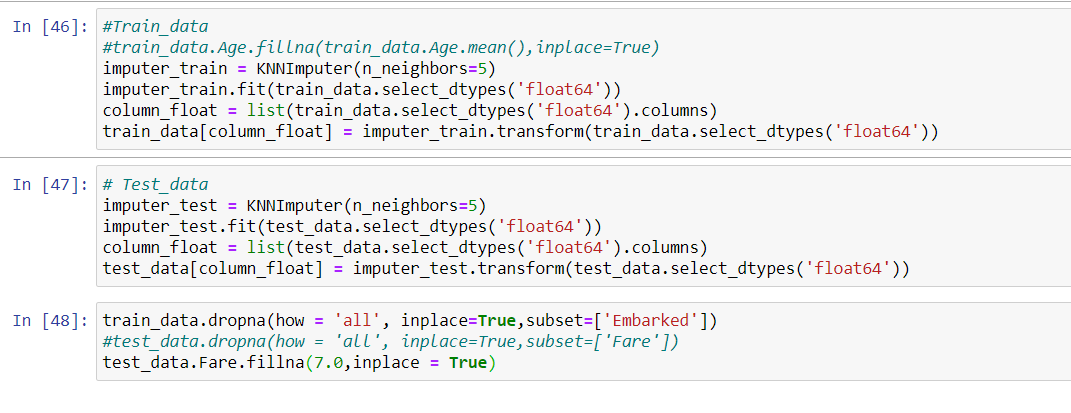


* Nhận xét thấy các cột “PassengerId, Name, Ticket” không có tác dụng trong việc phân lớp nên ta bỏ các cột này

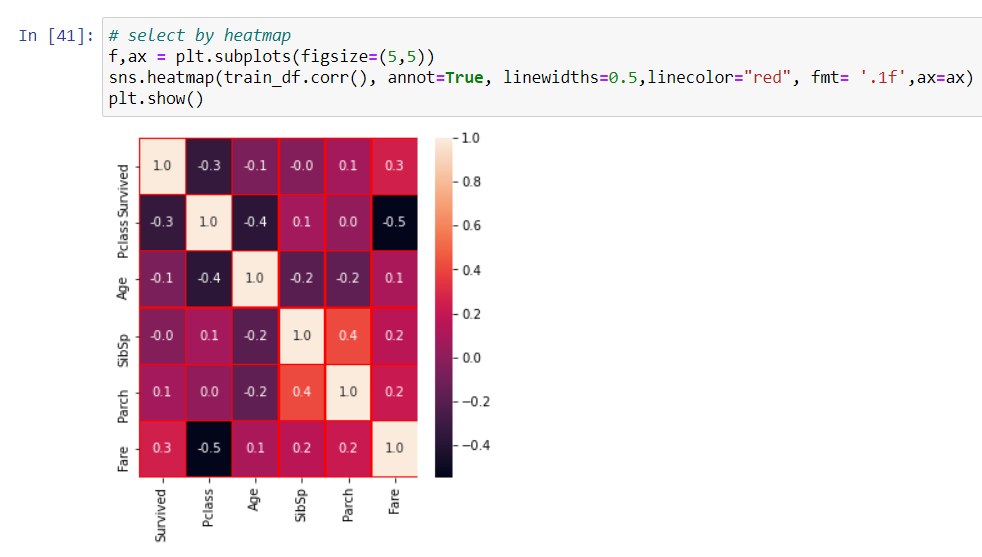
# Fill missing value

* Thống kê missing value ở các cột trong data\_train và data\_test

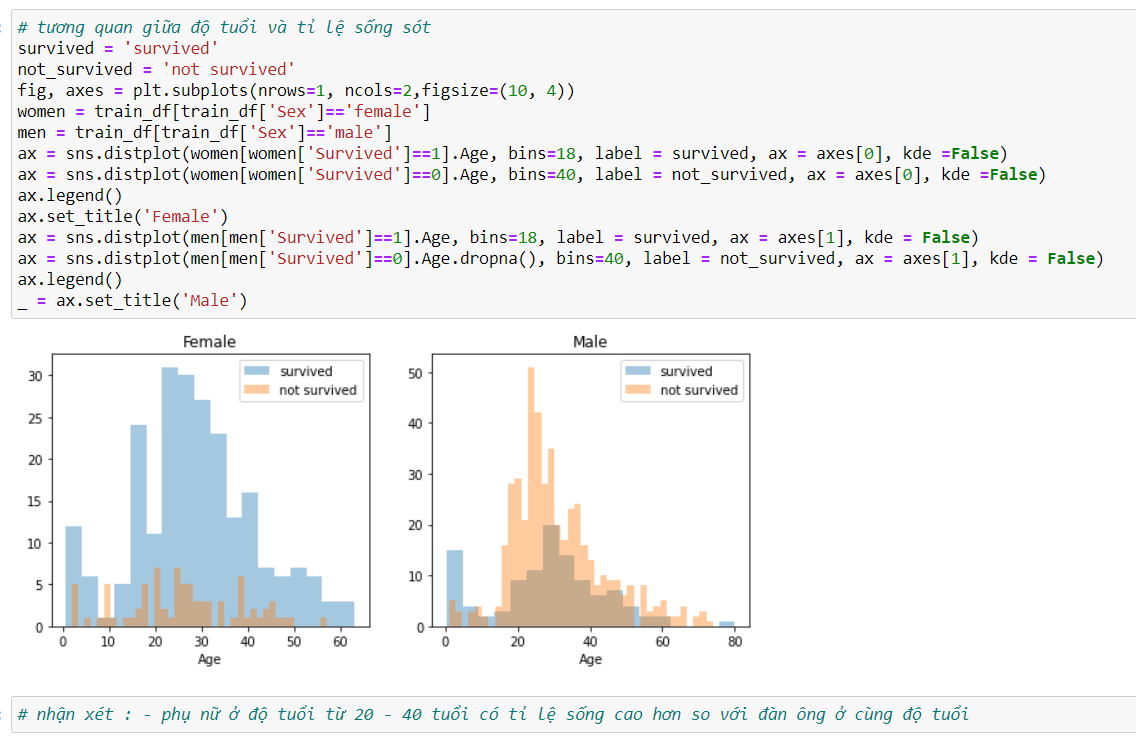
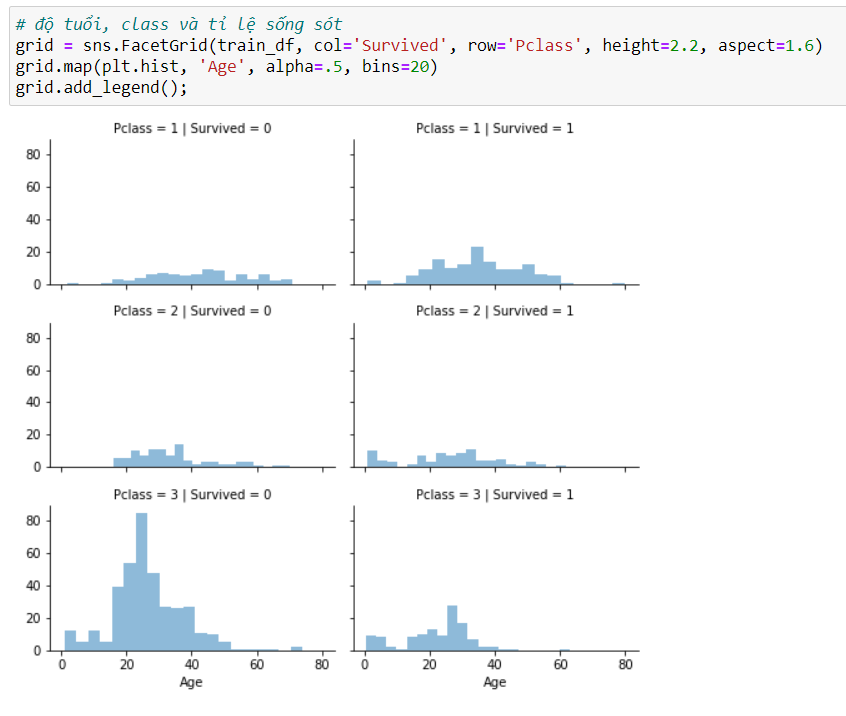


* Ta thấy cột Cabin ở cả hai bộ đều missing > 70% data => lựa chọn giải pháp bỏ cột này
* Các missing value ở cột Age ta dùng thuật toán KNN để dự đoán các dữ liệu bị thiếu
* Chuyển dữ liệu từ các cột từ String sang Float

# Feature selection

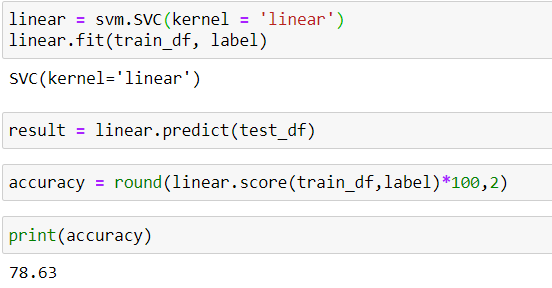
* Quan sát sự tương quan giữa các cột bằng heatmap
* ta thấy các columns : Age, Class, Fare có ảnh hưởng khá lớn đến Survived

……..tiếp tục nhận xét

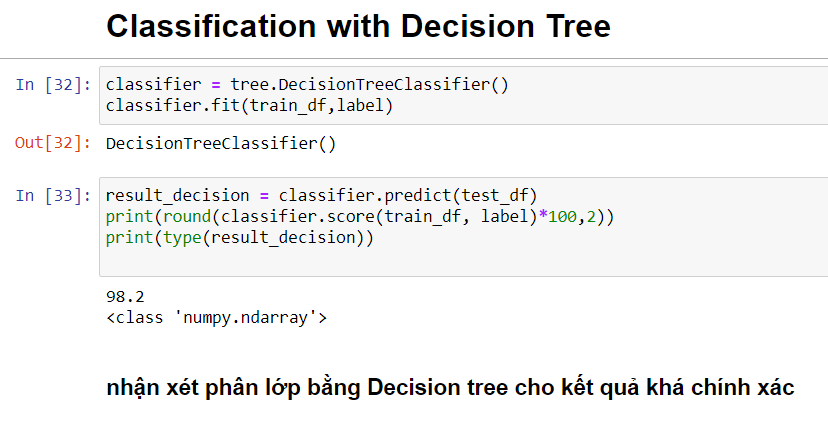
* Age và Survived ,Sex
* Sự ảnh hưởng của Class, Age đến survived

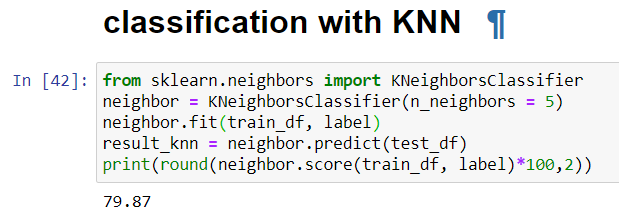
# Phân lớp bằng các phương pháp SVM, Decision Tree, KNN

1. SVM



* Tỉ lệ là 78,63
* Điểm trên kaggle là 0.72009

1. Decision Tree
2.  KNN



1. Random Forest

