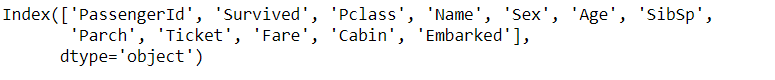
BÁO CÁO PHÂN LỚP VỚI TẬP DỮ LIỆU 1: TITANIC TASK

Họ tên: Phạm Văn Sang

Mã sinh viên: 19020416. Nhóm 3.

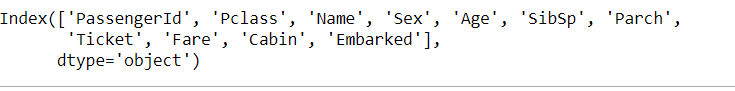
1. Data Analysis

Tập dữ liệu Titanic gồm 2 tập dữ liệu train.csv và test.csv.

* Tập dữ liệu train: 891 rows \* 12 columns, với các cột là các dữ liệu về hành khách (Passenger) gồm : 

Và Survied là cột chứa Label gồm các giá trị 0(Chết) hoặc 1(Sống).

* Tập dữ liệu test : 418 rows \* 11 columns, k có nhãn(Survived) :

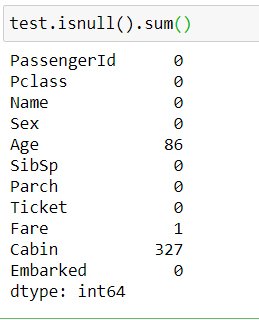


* Numerical Feature : Age(Continuous), Fare(Continuous), SibSp(Discrete), Parch(Discrete).
* Categorical Feature : Survived, Sex, Embarked, Pclass.
* Alphanumeric Feature : Ticket, Cabin.

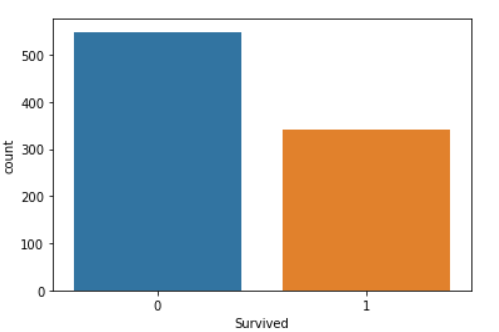
(Name : String)

* Mỗi tập dữ liệu train, test đều có chứa các missing values.

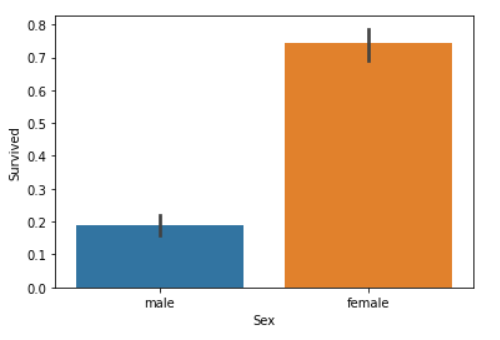
Train : Age : 177/891 missing values, Cabin: 687/891 missing values, Embarked: 2.891 Missing values

Test:

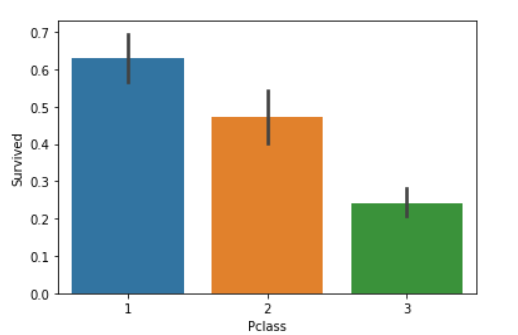
1. Data Visualization:

Sử dụng Matplotlib và Seaborn để dự đoán đặc điểm về nhãn của dữ liệu:

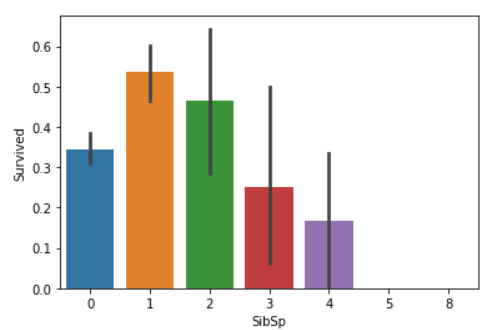
* Survived :
* Sex Feature: Nữ giới có tỷ lệ sống sót cao hơn nữ giới => cần thiết cho mô hình dự đoán.



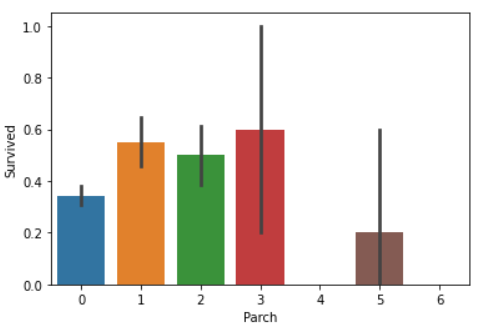
* Pclass Feature: Người với tầng lớp kinh tế cao hơn -> tỷ lệ sống sót cao hơn.

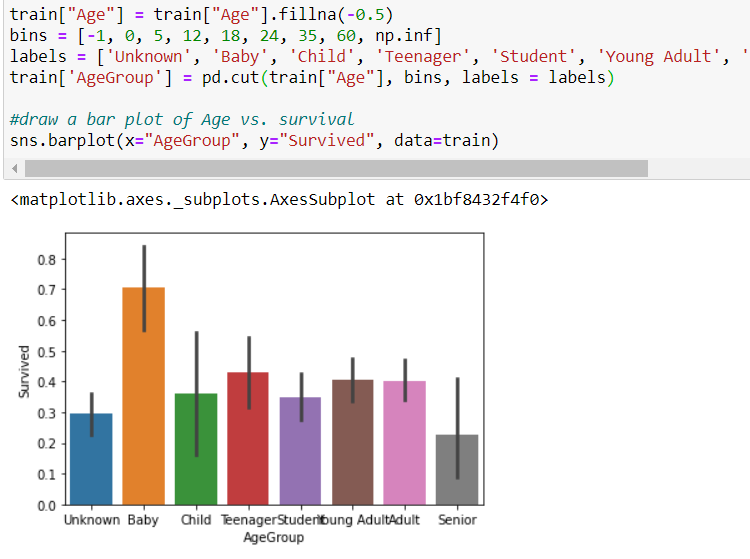
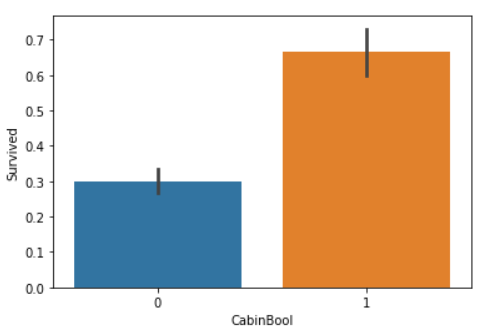


* SibSp Feature: Người có nhiều quan hệ họ hàng hoặc vợ(chồng) có ít cơ hội sống hơn, nhưng người kh có quan hệ nào có tỷ lệ sống ít hơn ng có 1-2 quan hệ.

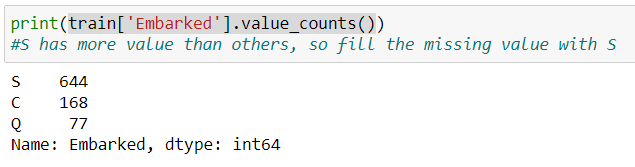
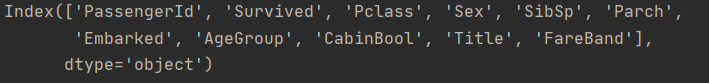


* Parch Feature: Người có ít hơn 4 bố mẹ hoặc con cái thì có tỷ lệ sống sót cao hơn ng có nhiều hơn 4.



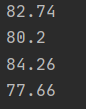
* Age Feature: Giá trị liên tục -> chia khoảng để dự đoán , các giá trị trống -> gán “Unknowns”. => Trẻ em có tỷ lệ cao hơn các độ tuổi còn lại.
* Cabin Feature: Người ở trong cabin có tỷ lệ sống cao hơn.

1. Cleaning Data(Trong source code):

* Embarked: Giá trị thiếu trong train đc fill = S.
* Cabin: Nhiều giá trị thiếu => Drop.
* Sex : Male-> 0, Female->1.
* Ticket : Drop.
* Embarked: “S” -> 3, “C” -> 2, “Q” -> 1.
* Age, Fare: Tách thành các nhóm và gán lại với các dữ liệu số.
* Các đặc trưng còn lại sau khi clean :

1. Model:

Để tránh overfitting cho tập dữ liệu train, tách tập dữ liệu train thành 2 phần: 78% để huấn luyện và 22% để đánh giá. Sử dụng Support Vector Machine, Decision Tree, Random Forest, K-NearestNeighbors để huấn luyện mô hình

 Độ chính xác của mô hình khi dự đoán cho tập dữ liệu còn lại, lần lượt : SVM, DT, RS và kNN: