Họ tên: Lưu Hoài Linh

MSSV: 17021284

Nhóm 1

Titanic task - report

Github:

1. Trong folder NLP Task gồm những files sau:

|  |  |
| --- | --- |
| * Titanic Task.ipynb: | code  code |
| * titanictask.py: |
| * titanic\_report.docx: | report |
| * train.txt: | input |
| * test.txt: | input |
| * submission.csv: | output: label của dữ liệu file testing.txt |

2. Xây dựng mô hình phân lớp gồm 4 bước

* Pre-processing data

+ Xử lý những giá trị bị thiếu

+ Quyết định xem nên drop những thuộc tính nào

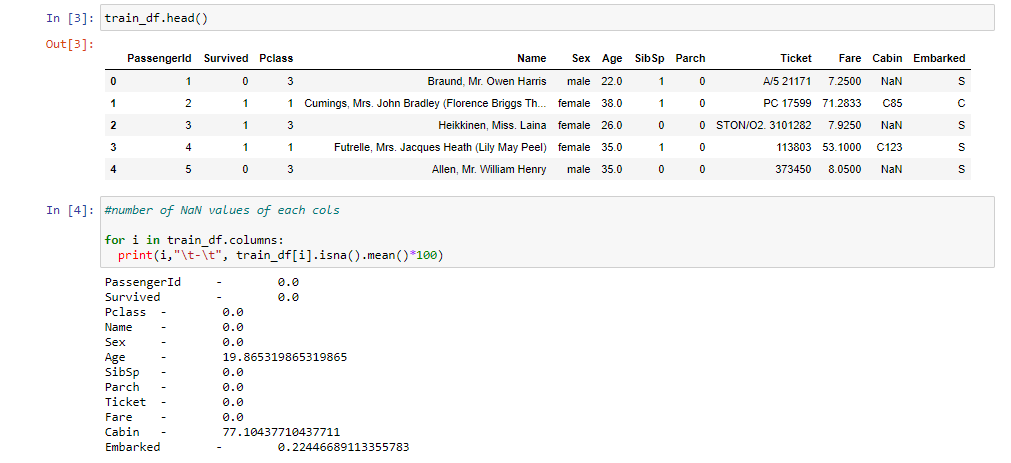
* Training

+ Sử dụng Logistic Regression để train model

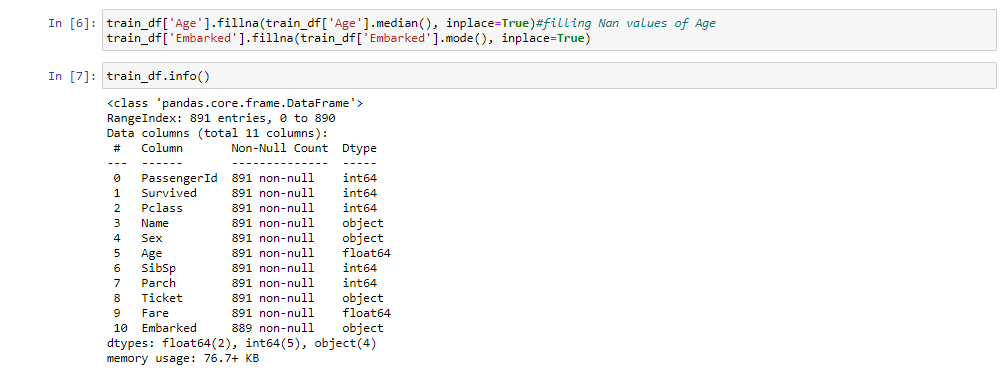
* Test
* Accuracy
* Classify

2.1: Data pre-processing:

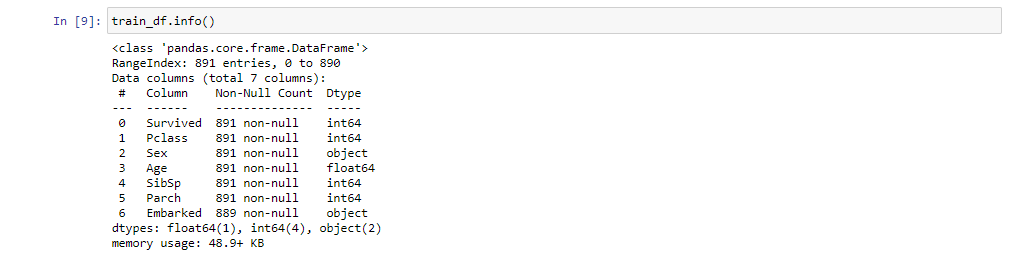
* Load file train.csv và tính tổng những giá trị bị thiếu:



* Xử lý các giá trị bị thiếu ở cột *Age* và *Embarked*



* Drop 3 cột: *PassengerID, Fare, Ticket, Name, Cabin* vì chứa nhiều giá trị *Na* và không ảnh hưởng đến việc phân lớp. Dữ liệu lúc này được thống kê:

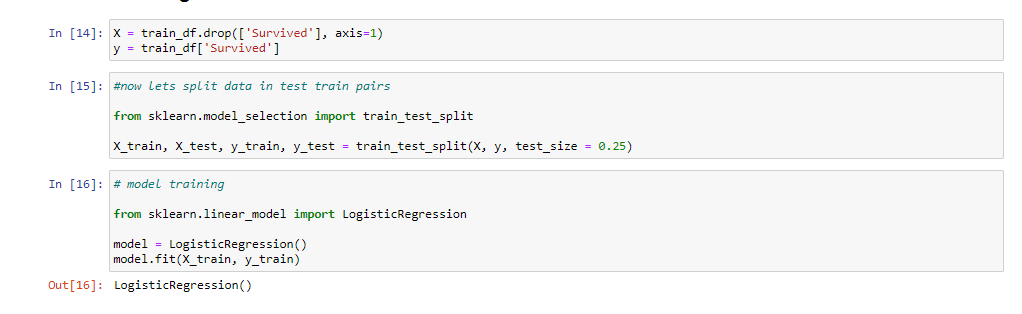


* Chuyển giá trị *Sex* và *Embarked* từ string sang number



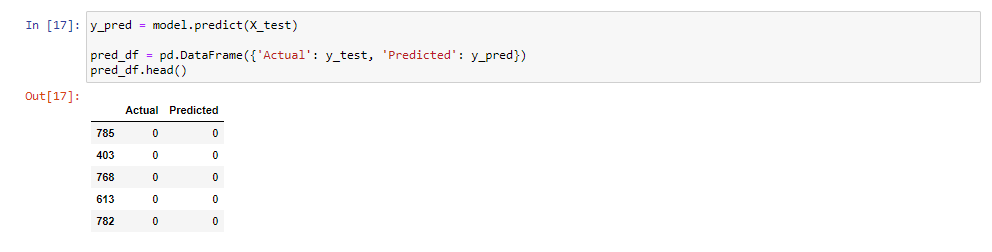
2.2. Training

* Chia tập training data thành 2 phần: trainning và test để kiểm tra độ chính xác của mô hình (*X\_train, X\_ test, y\_train, y\_test*) với test\_size = 0.25



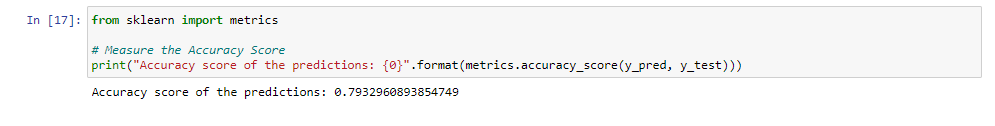
2.3. Test

* Kết quả test giữa kết quả dự đoán và kết quả thực tế



2.4. Accuracy

* Độ chính xác đo được, xấp xỉ = 0.8



2.5. Classify (predict)

* Kết quả phân lớp được lưu vào file *submission.csv*
* File có 418 dòng (418 khách) và 2 cột (tên hành khách và kết quả phân lớp 1: sống hoặc 0: chết)



3. Kết quả

* Các label cần dự đoán nằm trong file: *submission.csv*