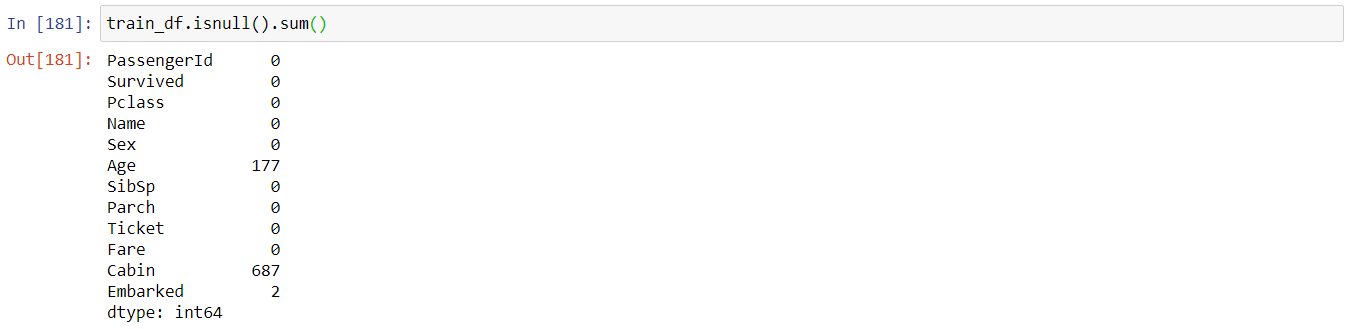
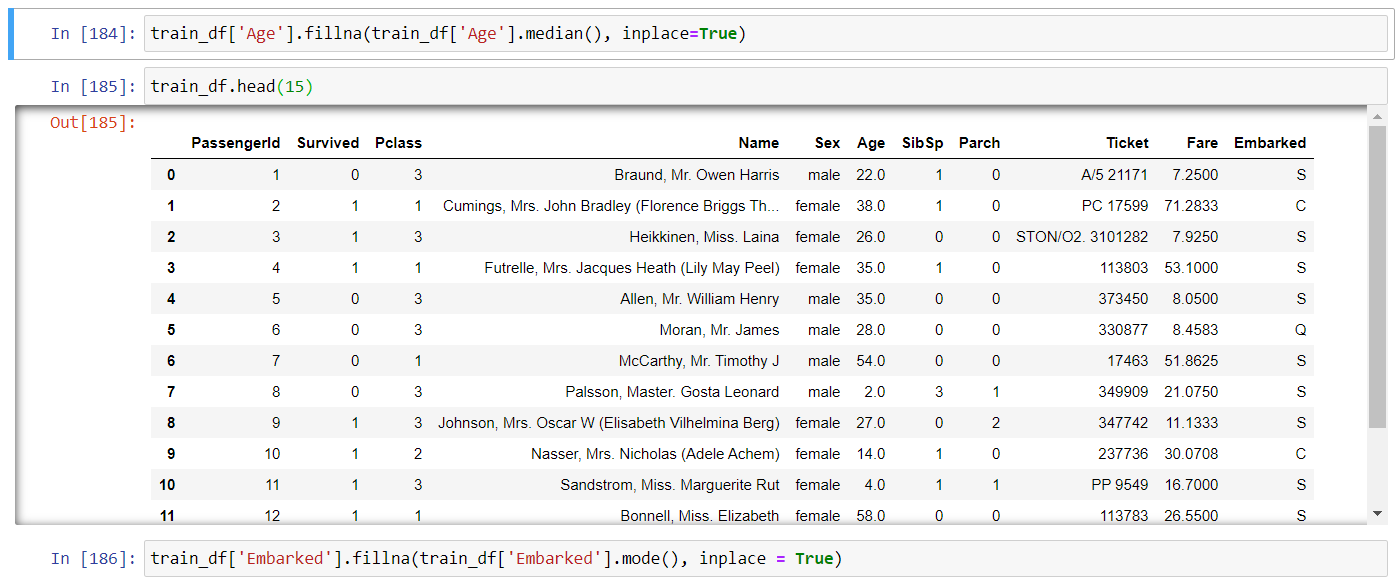
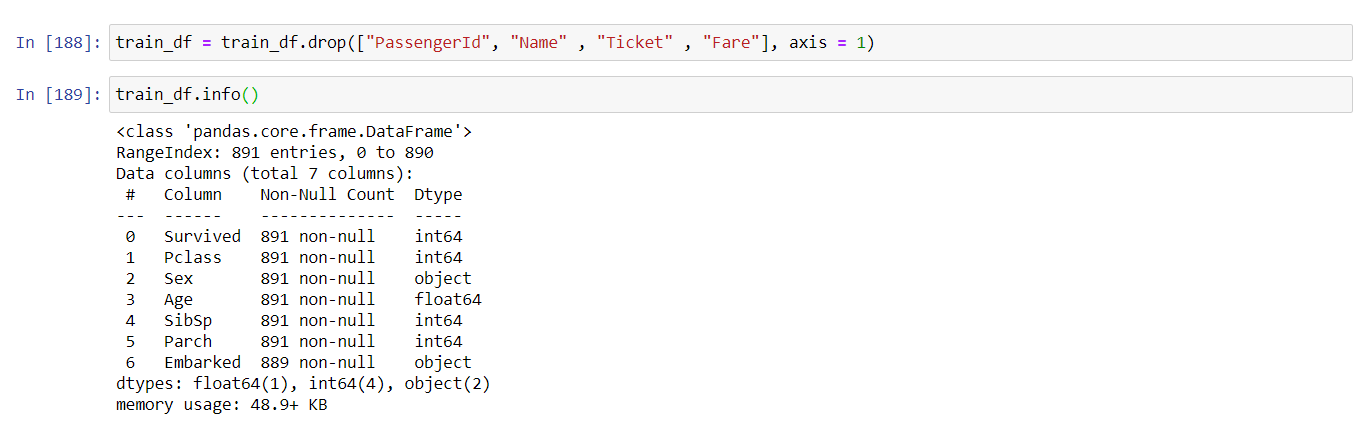
1. **Khảo sát dữ liệu, thay thế các giá trị null bằng các giá trị tùy chọn cũng như loại bỏ các cột không cần thiết.**



Từ tập dữ liệu train ta nhận thấy thuộc tính Age có 177 giá trị null, thuộc tính Cabin có 687 giá trị null và thuộc tính Embarked có 2 giá trị null. Do Cabin có quá nhiều giá trị null nên ta loại bỏ thuộc tính Cabin.



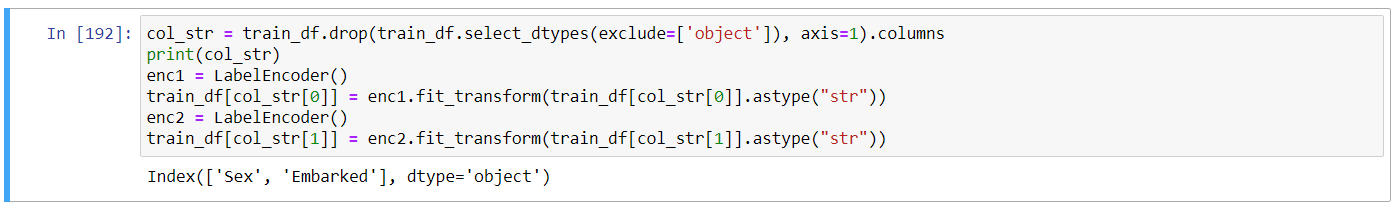
Các giá trị null của thuộc tính Age được thay bằng giá tuổi ở giữa các tuổi còn giá trị null của thuộc tính Embraked được thay bằng giá trị xuất hiên nhiều nhất của Embraked.

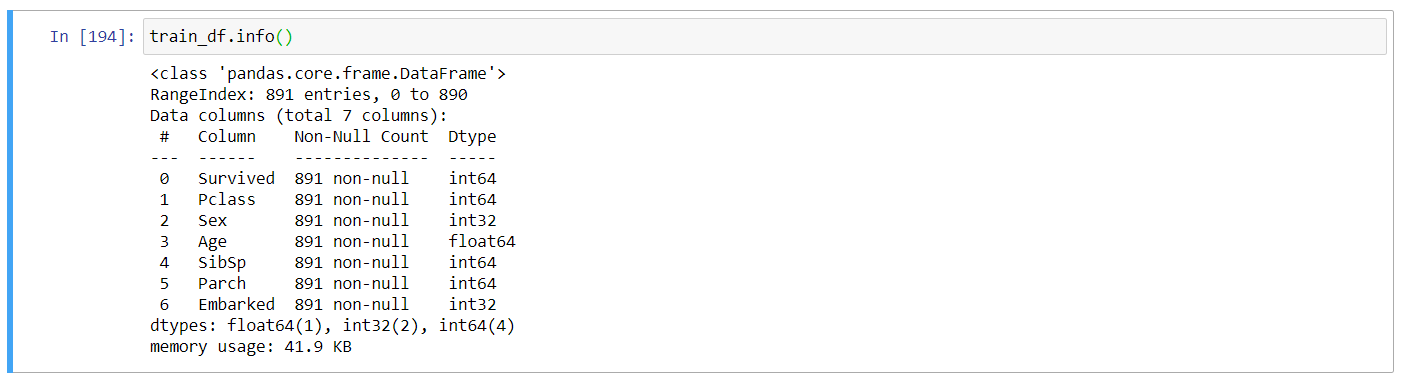


*Drop 4 cột thuộc tính PassengerID, Fare, Ticket, Name* do không ảnh hưởng nhiều đến việc phân lớp.

1. **Tiến hành chuyển dạng chuỗi sang dạng number.**

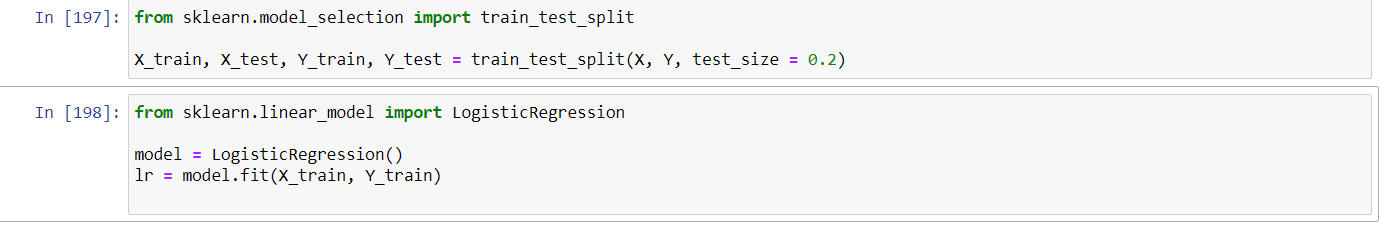
Ở đây có 2 cột khác number là cột Sex và cột Embarked là dạng object do đó ta cần chuyển sang dạng number.





1. **Training**

* Sau khi làm sạch dữ liệu cũng như chuyển đổi dữ liệu sang dạng số ta bắt đầu vào việc 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. Và thuật toán được chọn ở đây là hồi quy Logistic phù hợp với các dữ liệu kiểu vượt qua / thất bại, thắng / thua, sống / chết hoặc khỏe / ốm.



1. **Kiểm tra độ chính xác của hồi quy Logistic**



Đô chính xác lên đến 82%. Khá là tốt và chỉ có sự chênh lệch tầm 18%.

1. **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)