**Bài tập thực hành 02:**

Bài toán hồi quy

# **Mục tiêu cần đạt**:

Biết cách cài đặt/ứng dụng luồng xử lý dữ liệu chung cho bài toán hồi quy bao gồm

* Load dữ liệu từ thư viện python hoặc từ csv
* Phân chia dữ liệu, đưa dữ liệu vào giải thuật học
* Đánh giá/nhận xét kết quả học trên các tập dữ liệu

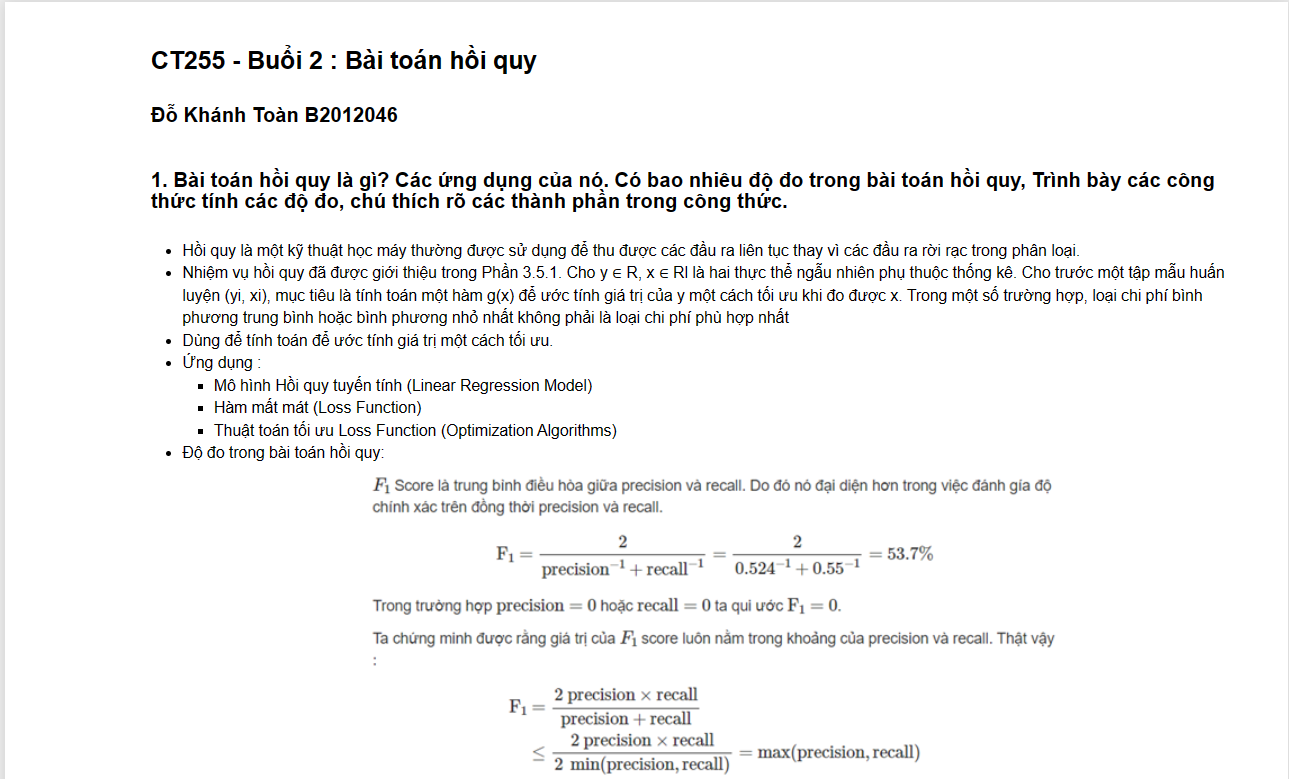
Các bạn thực hiện theo hướng dẫn trong giáo trình từ **trang 69 đến 76**, và trả lời các câu hỏi yêu cầu thực hành, có thể tham khảo thêm các link từ kaggle, github hoặc các nguồn khác để hỗ trợ trả lời câu hỏi (nhớ trích dẫn).

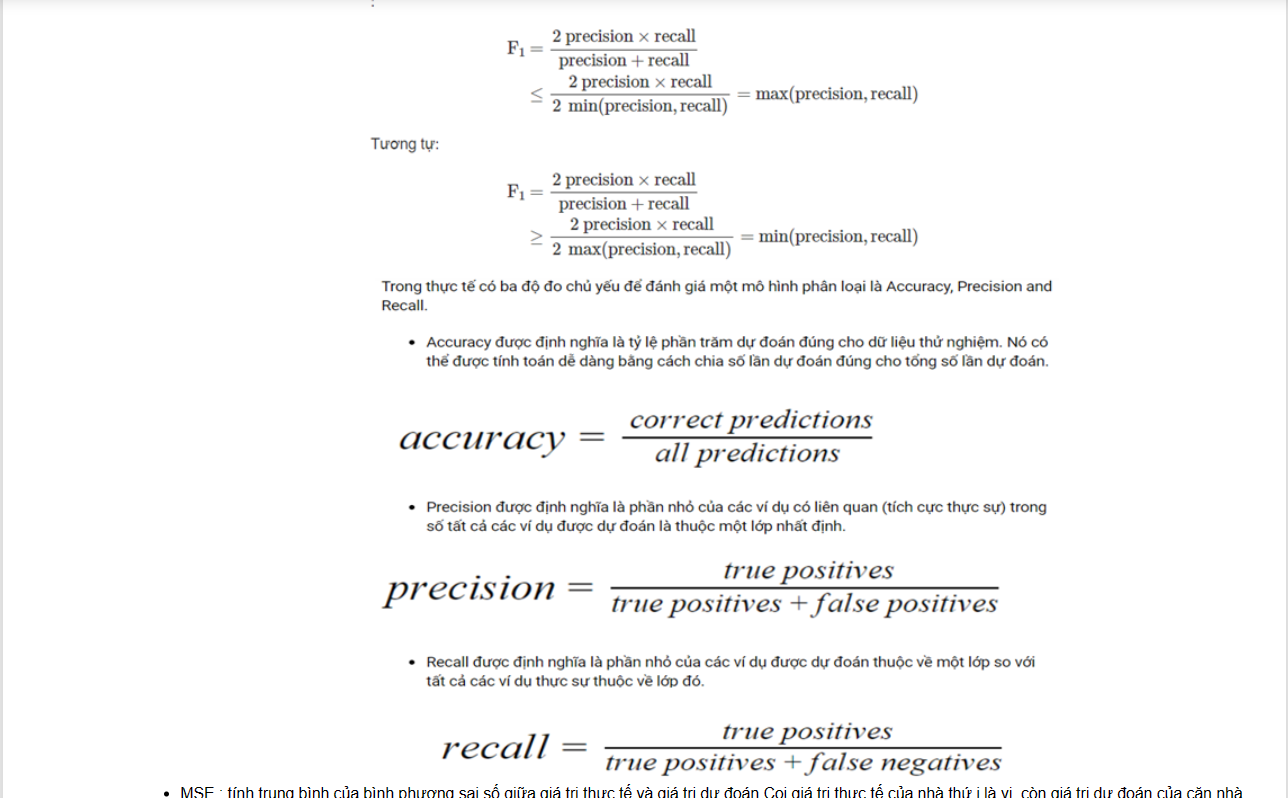
# Yêu cầu bài thực hành:

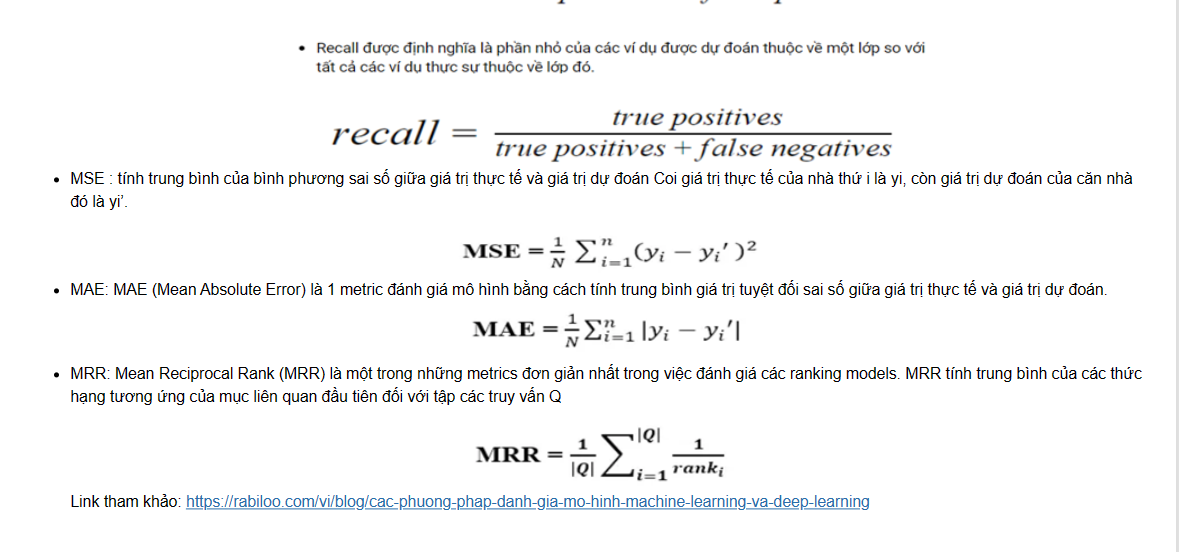
* *Với file word: Với các câu hỏi về lệnh các bạn trả lời bằng text: copy lệnh (dạng text) vào bài word trả lời và kèm hình ảnh minh họa. Tất cả câu trả lời phải có hình minh họa minh chứng kết quả*
* *Với lệnh trong Jupyter: ở đầu mỗi cell các bạn tạo ghi chú để rõ phần trả lời của câu nào*

1. Bài toán hồi quy là gì[[1]](#footnote-0)? Các ứng dụng của nó. Có bao nhiêu độ đo trong bài toán hồi quy, Trình bày các công thức tính các độ đo, chú thích rõ các thành phần trong công thức.

**Bài làm**

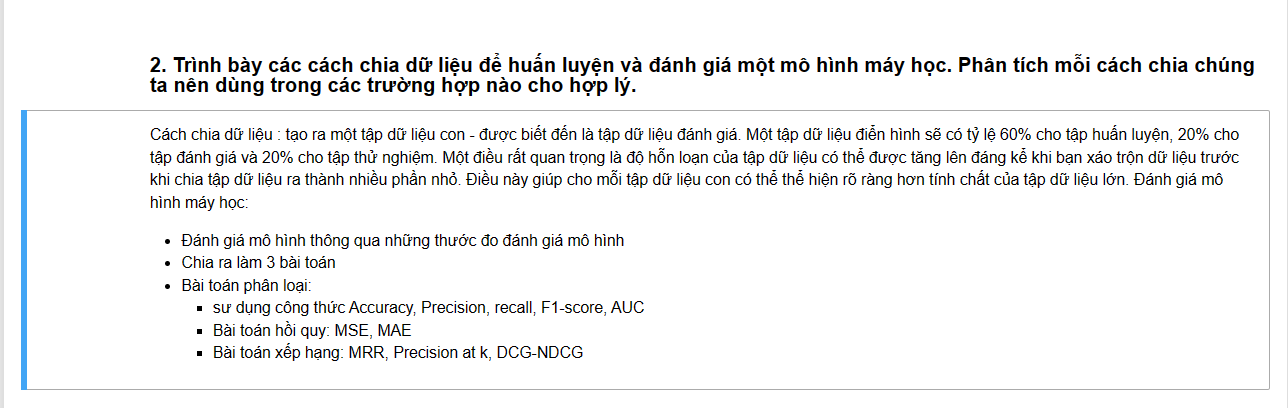






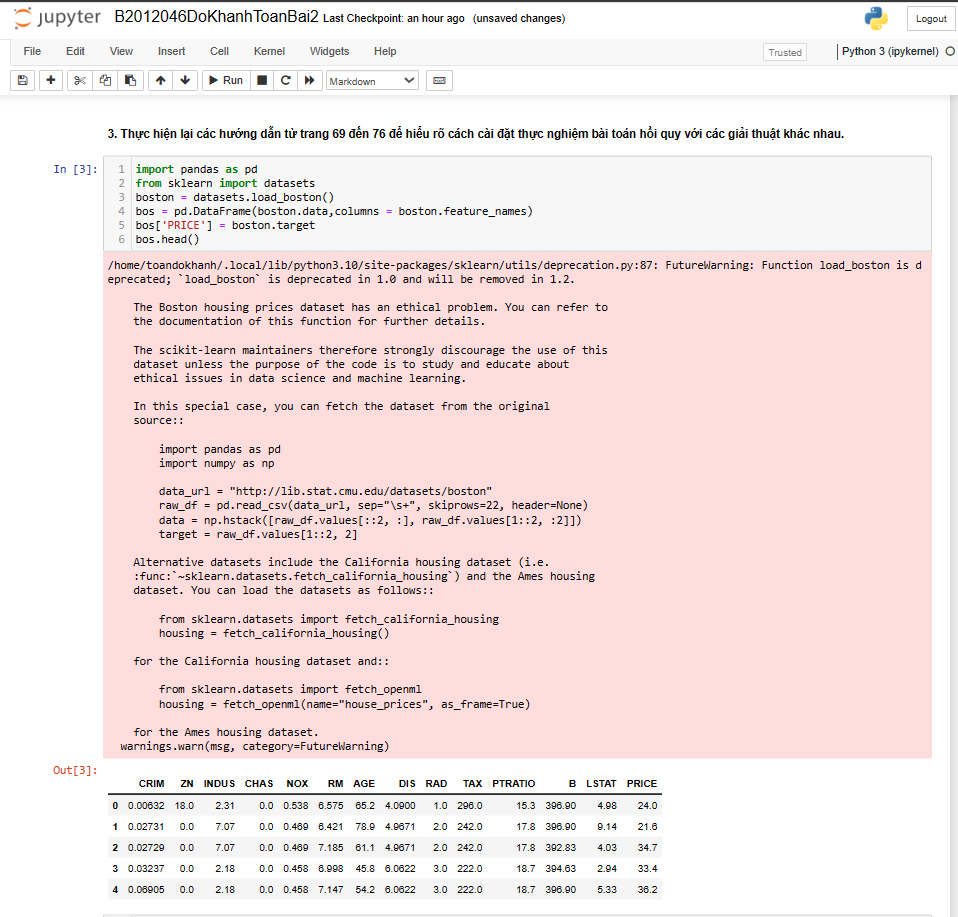
1. Trình bày các cách chia dữ liệu để huấn luyện và đánh giá một mô hình máy học. Phân tích mỗi cách chia chúng ta nên dùng trong các trường hợp nào cho hợp lý.

**Bài làm**

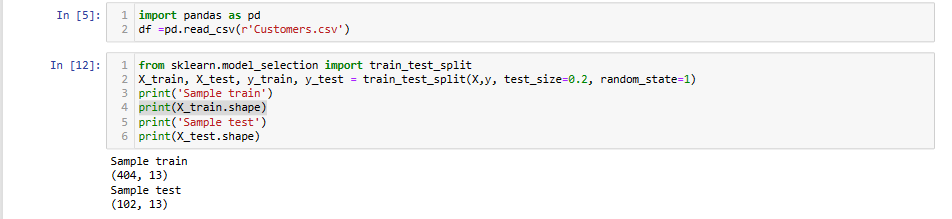


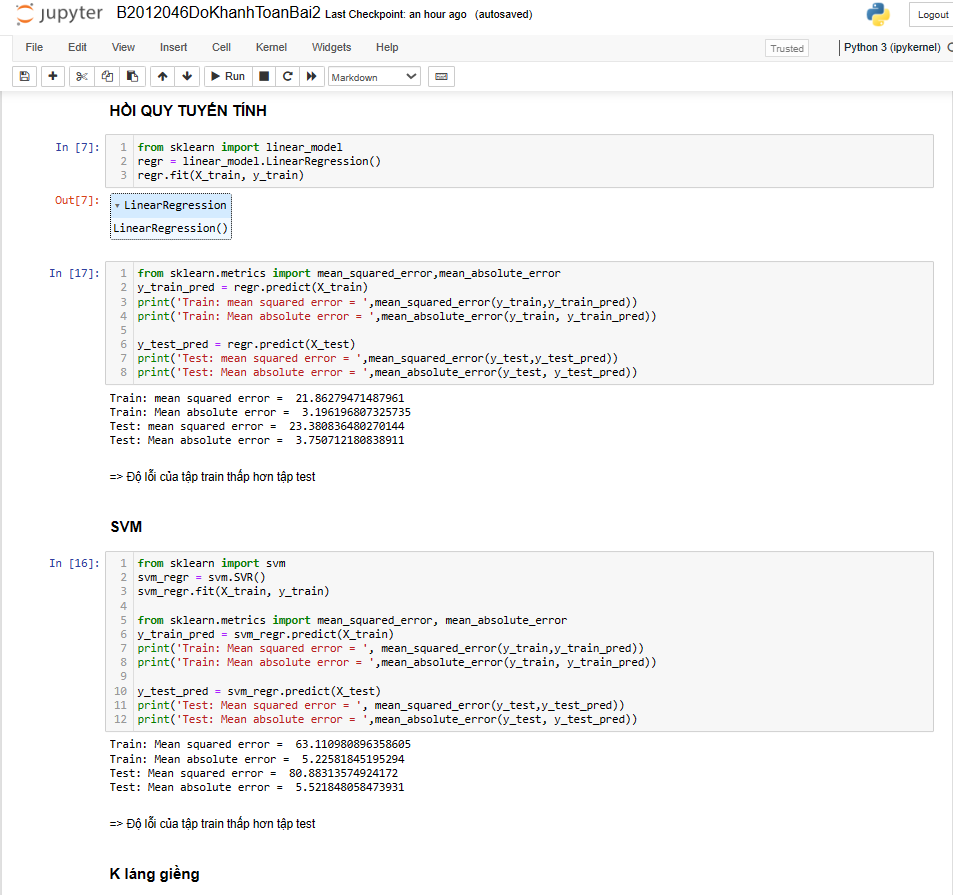
1. Thực hiện lại các hướng dẫn từ trang 69 đến 76 để hiểu rõ cách cài đặt thực nghiệm bài toán hồi quy với các giải thuật khác nhau.

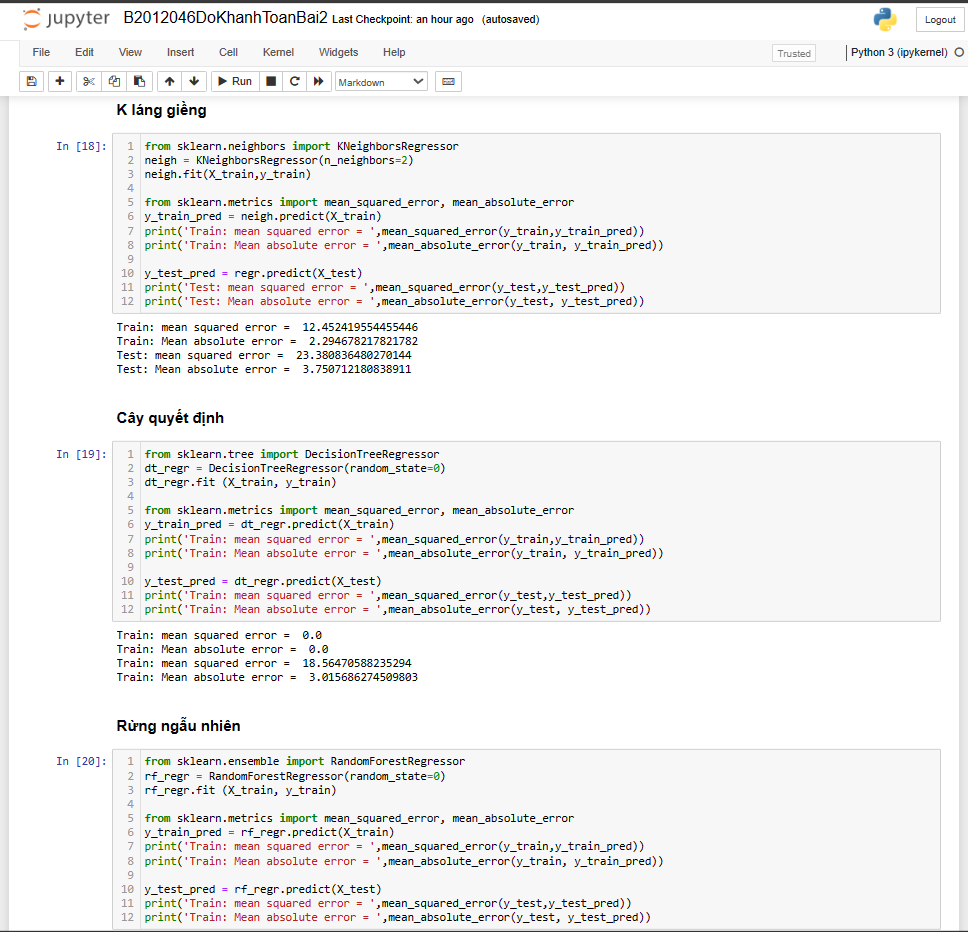
**Bài làm**

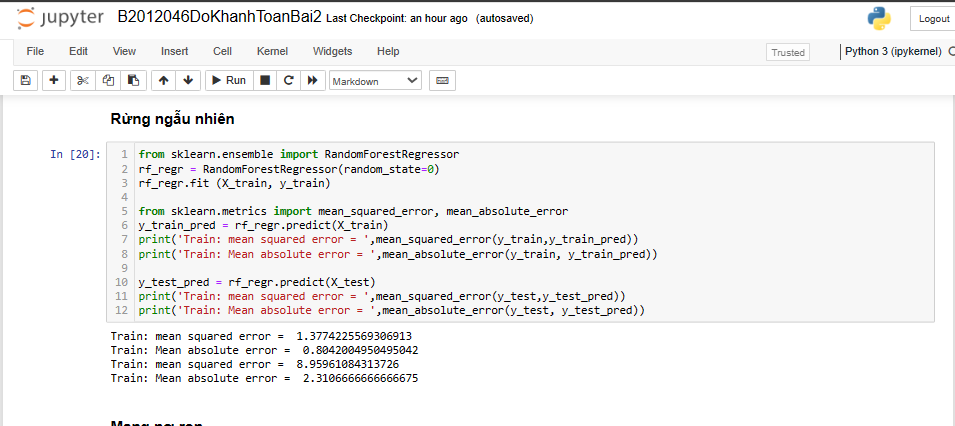


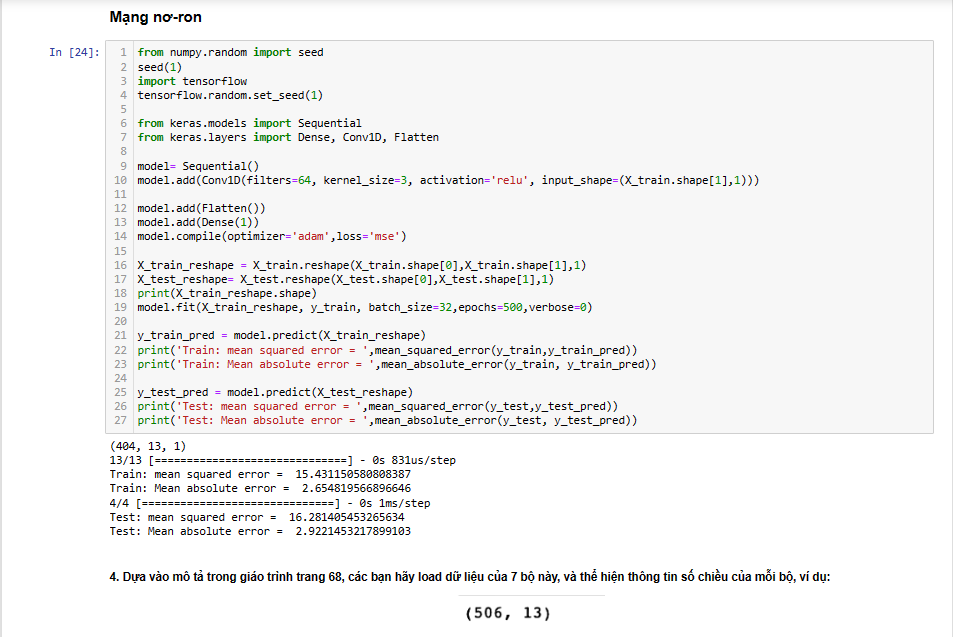












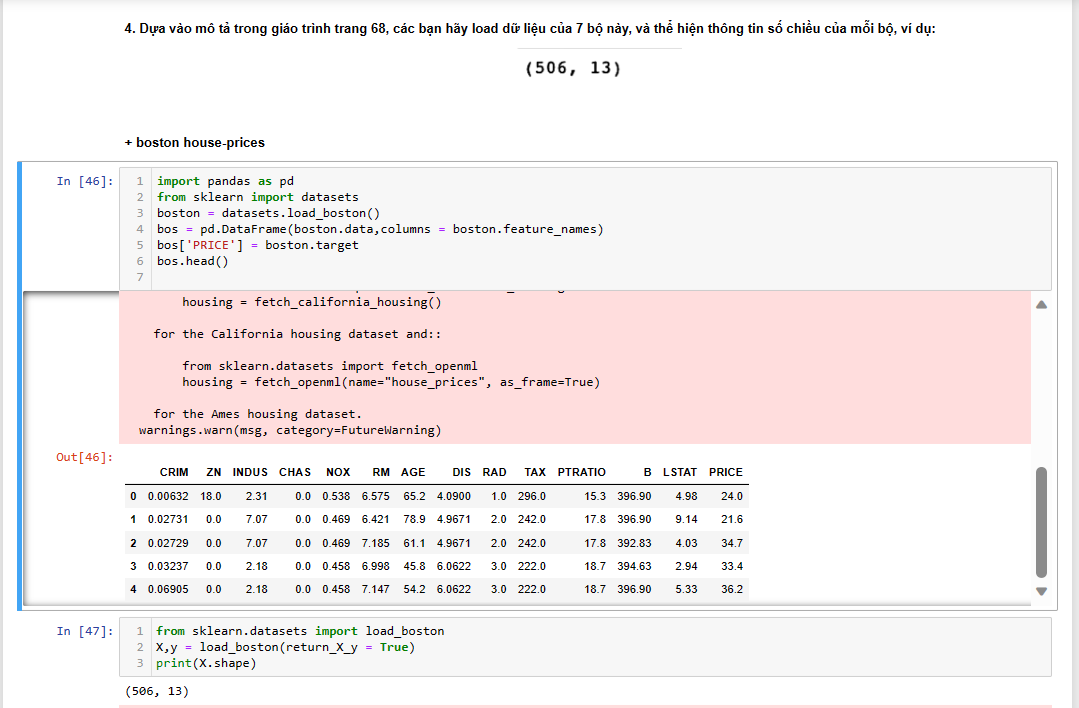
1. Dựa vào mô tả trong giáo trình trang 68, các bạn hãy load dữ liệu của 7 bộ này, và thể hiện thông tin số chiều của mỗi bộ, ví dụ:



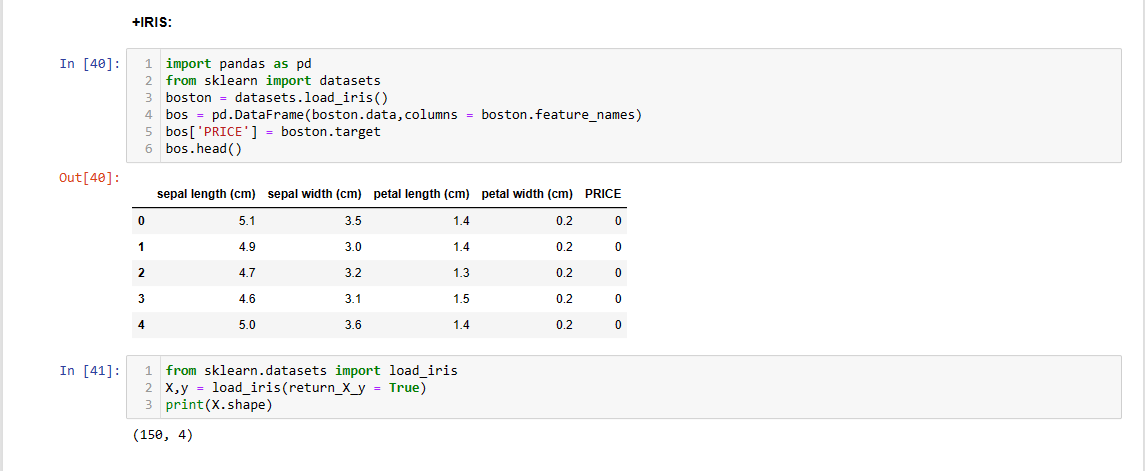
Các lệnh liên quan mỗi bộ dữ liệu các bạn đặt cùng 1 cell trong *Jupyter*

**Bài làm**

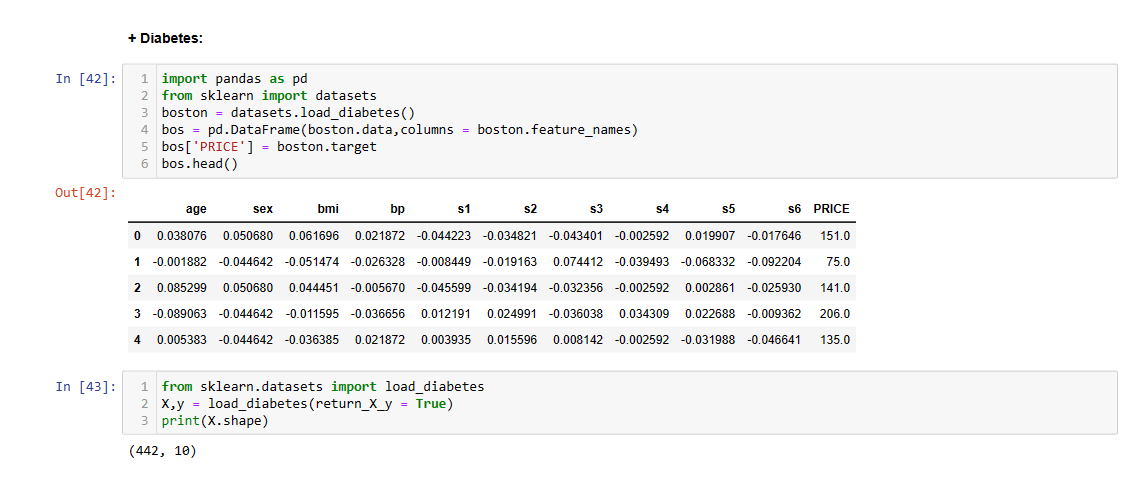
#### **+ boston house-prices**



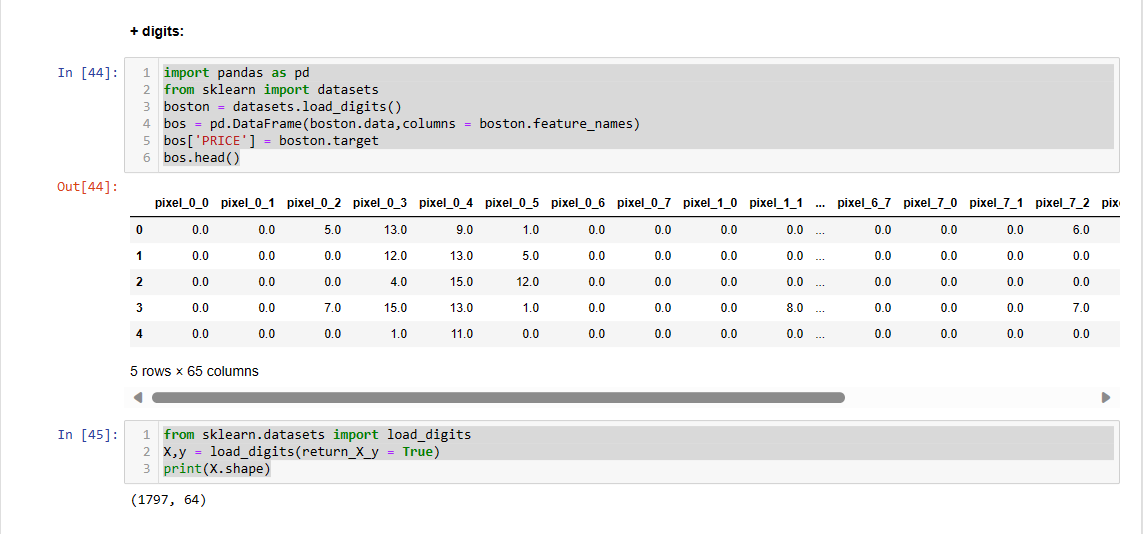
#### **+IRIS**



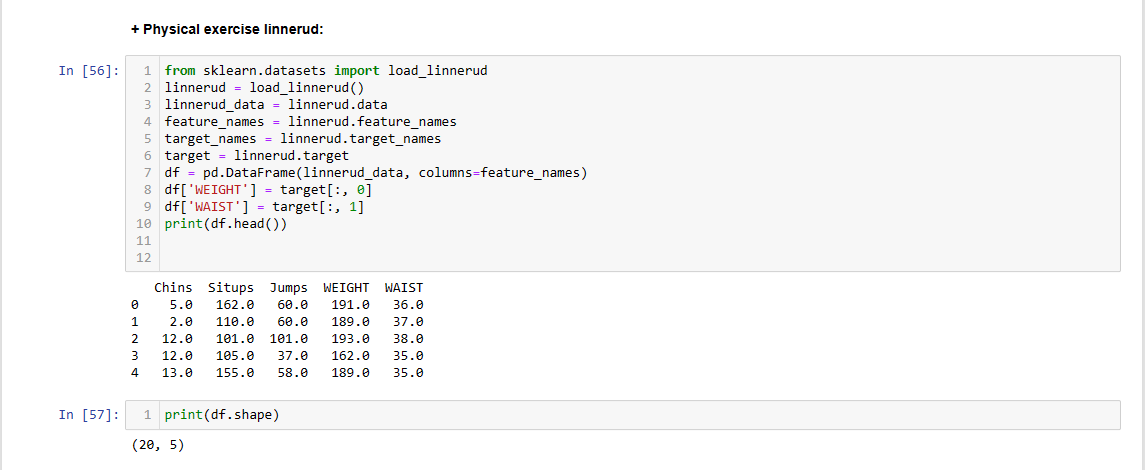
#### **+ Diabetes**



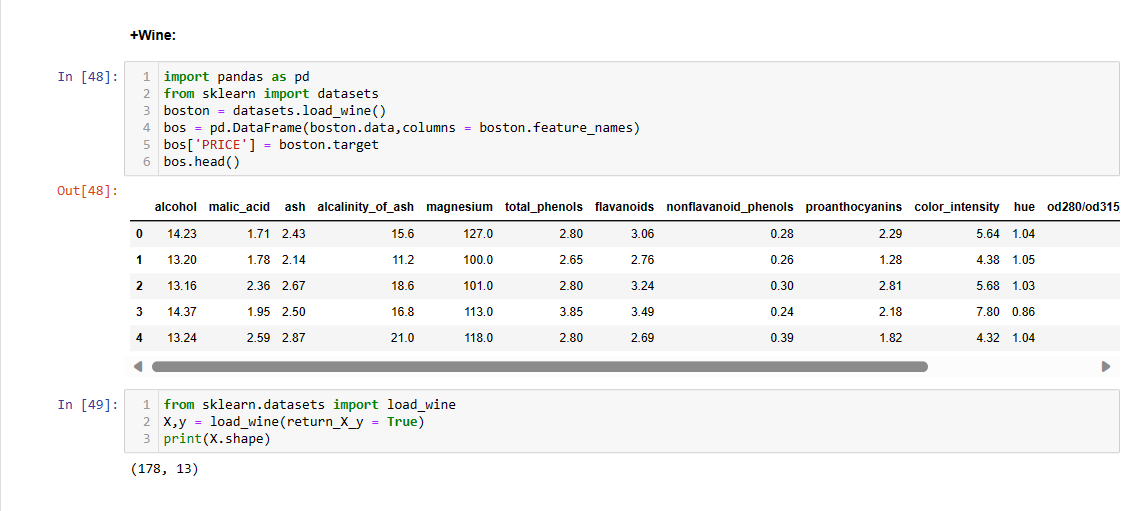
#### **+ digits**



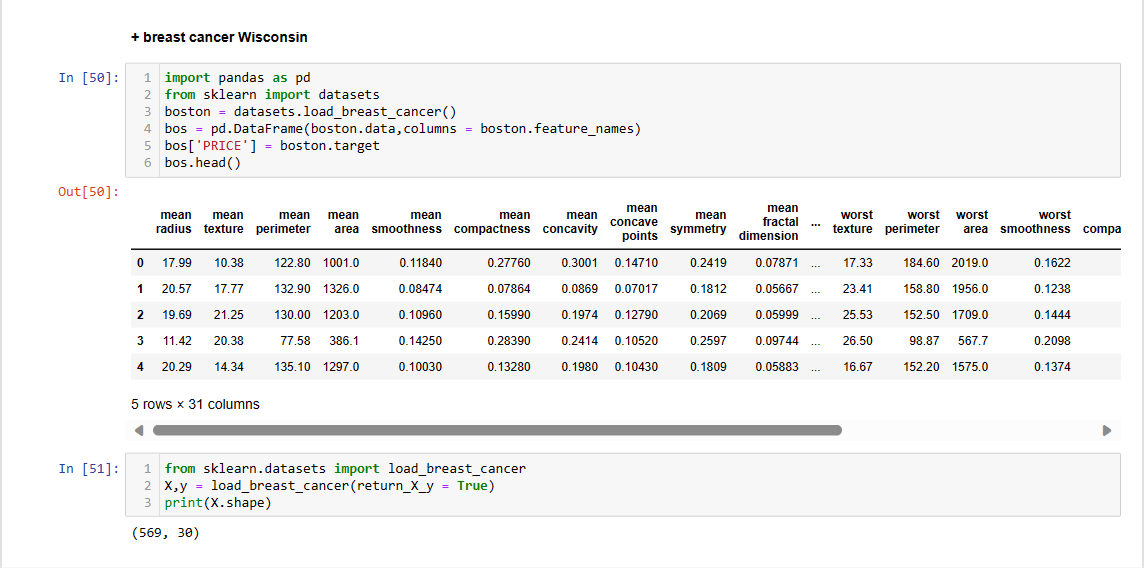
#### **+ Physical exercise linnerud**



#### **+Wine**

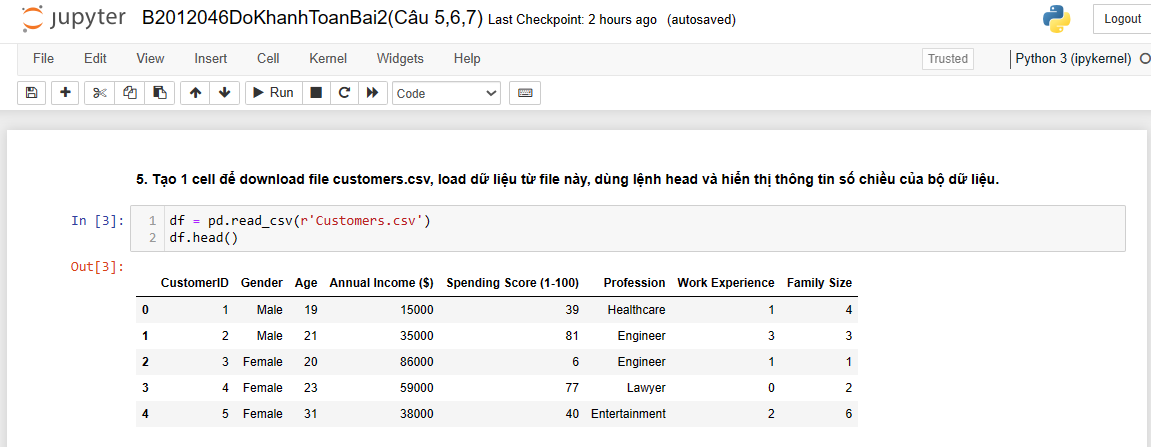


#### **+ breast cancer Wisconsin**



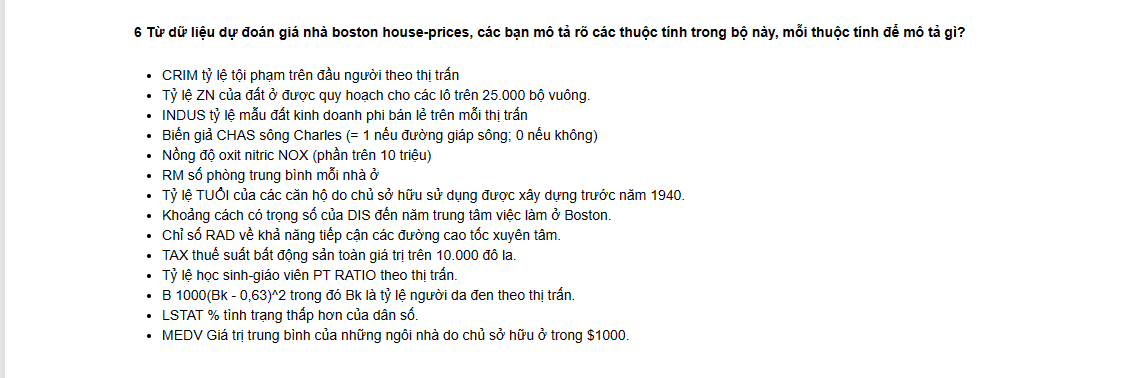
1. Tạo 1 cell để download file customers.csv, load dữ liệu từ file này, dùng lệnh head và hiển thị thông tin số chiều của bộ dữ liệu.

**Bài làm**



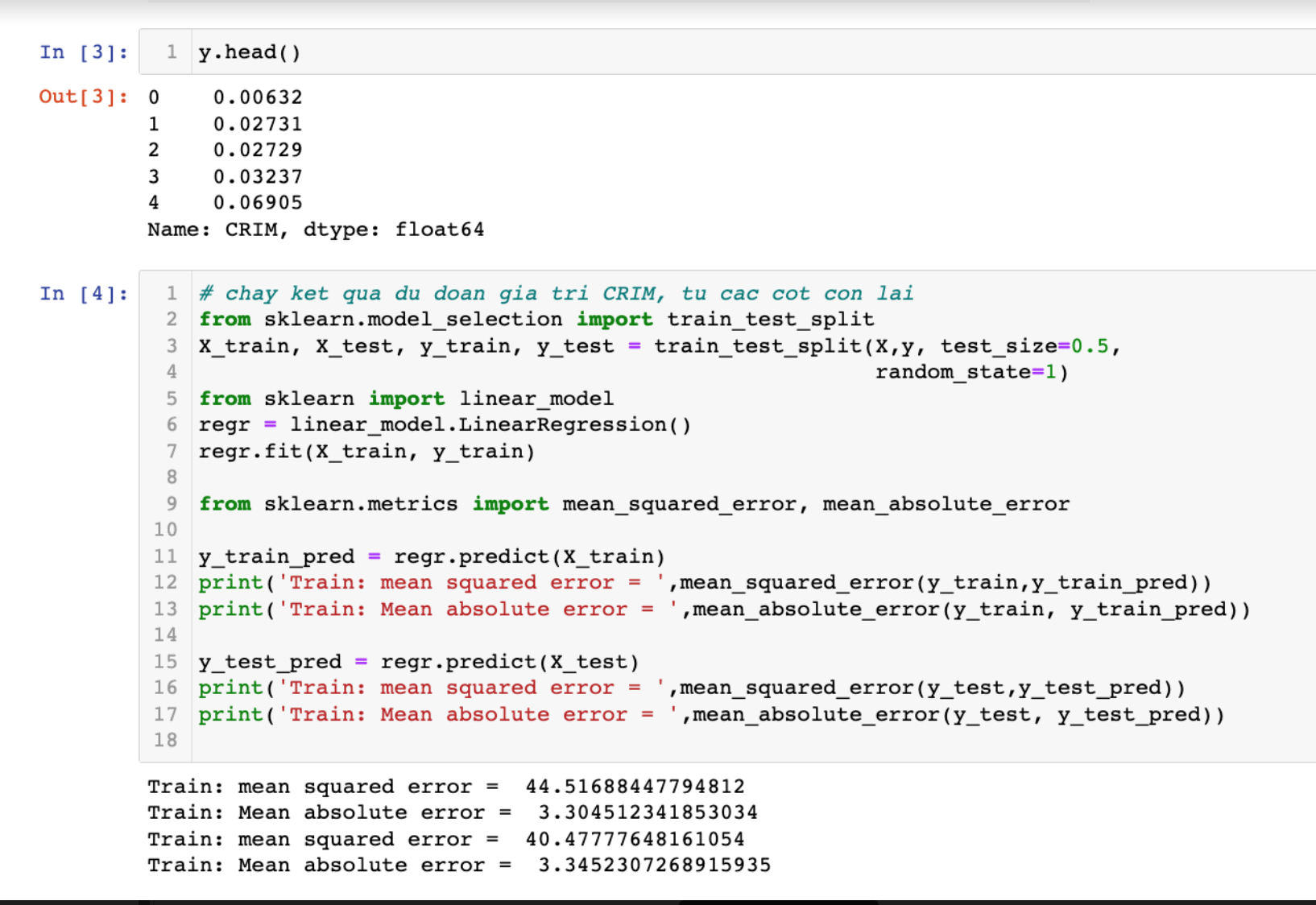
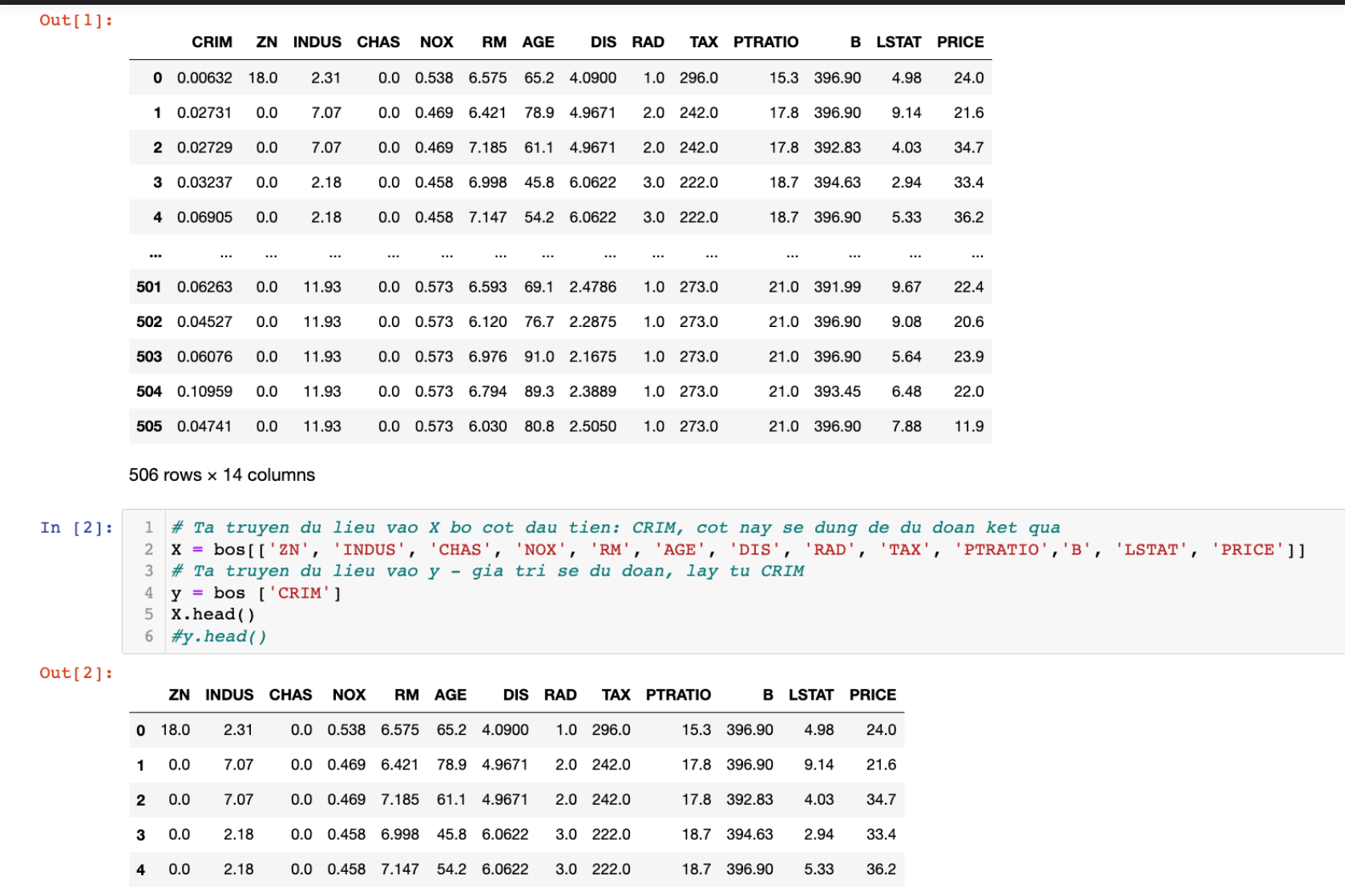
1. Từ dữ liệu dự đoán giá nhà boston house-prices, các bạn mô tả rõ các thuộc tính trong bộ này[[2]](#footnote-1), mỗi thuộc tính để mô tả gì?

**Bài làm**



1. Các bạn thay thế giá trị dự đoán trong boston house-price thành 1 trong 12 thuộc tính còn lại và chạy bài toán hồi quy với Rừng ngẫu nhiên với tỷ lệ tập huấn luyện là 50%, tập kiểm tra là 50%. Thực hiện lặp lại cho đến khi chạy xong hết 12 thuộc tính (trừ thuộc tính giá nhà, **PRICE**, đã trình bày trong giáo trình). Khối lệnh trình bày dự đoán của 1 thuộc tính trình bày trong 1 cell (có tổng cộng 12 cell cho câu này).

*Gợi ý: Các bạn đổi cột để thực hiện dự đoán đối với dữ liệu data frame cho dataset boston*

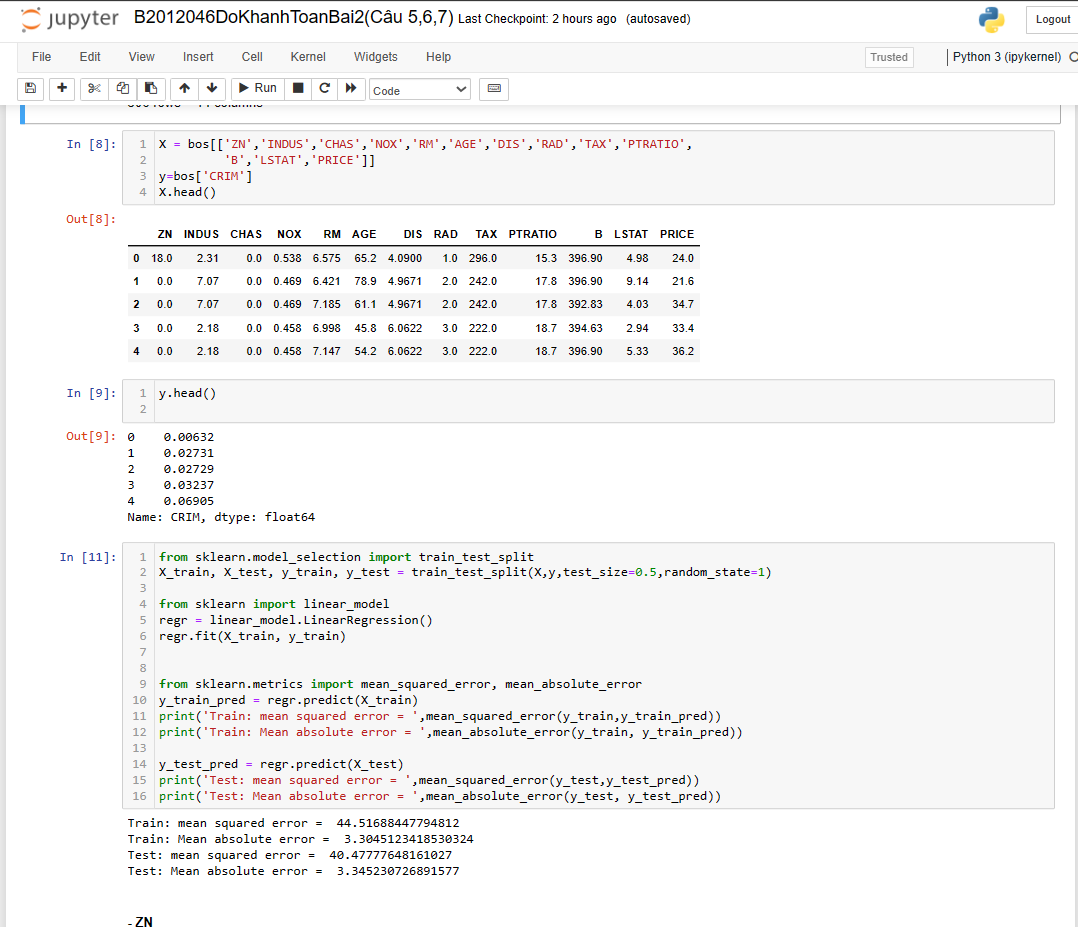


=>

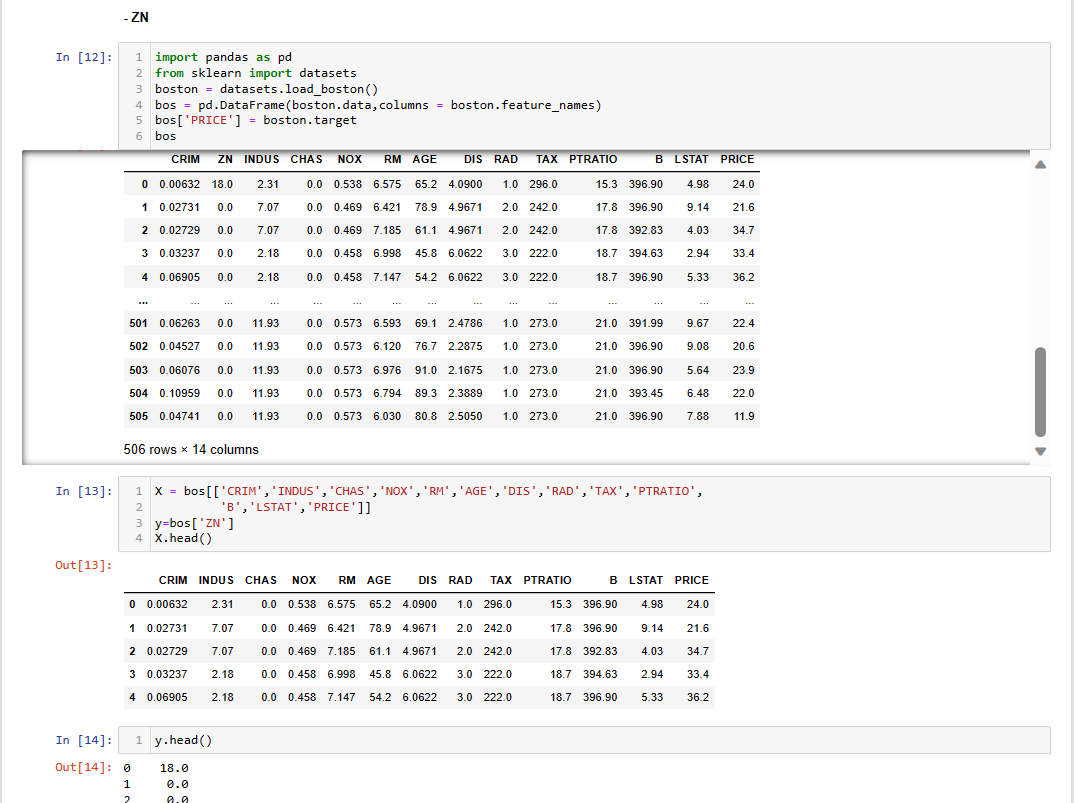
**Bài làm**

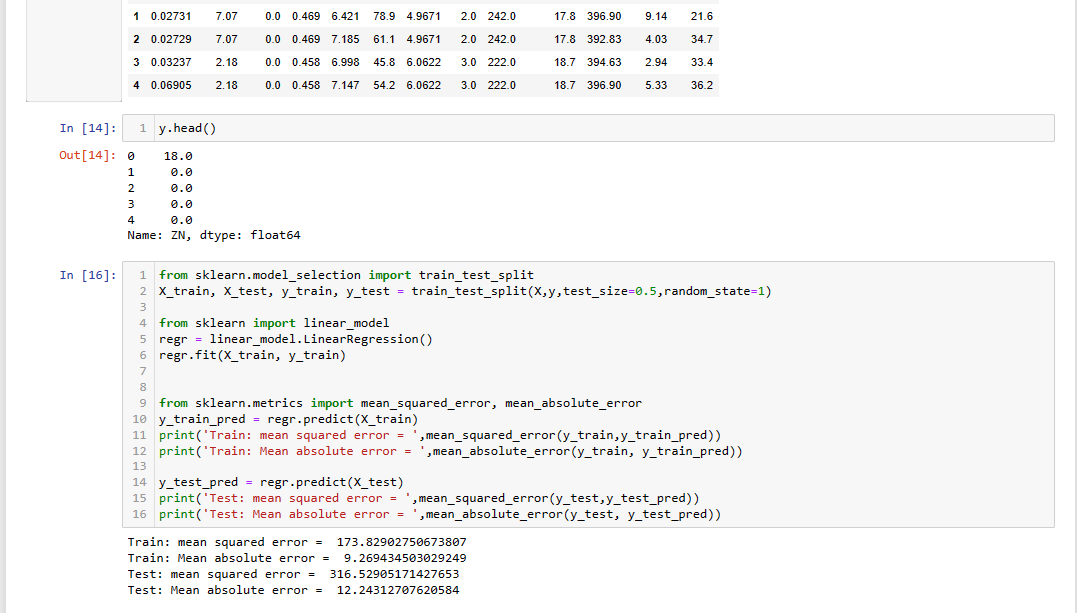
**- CRIM**

****

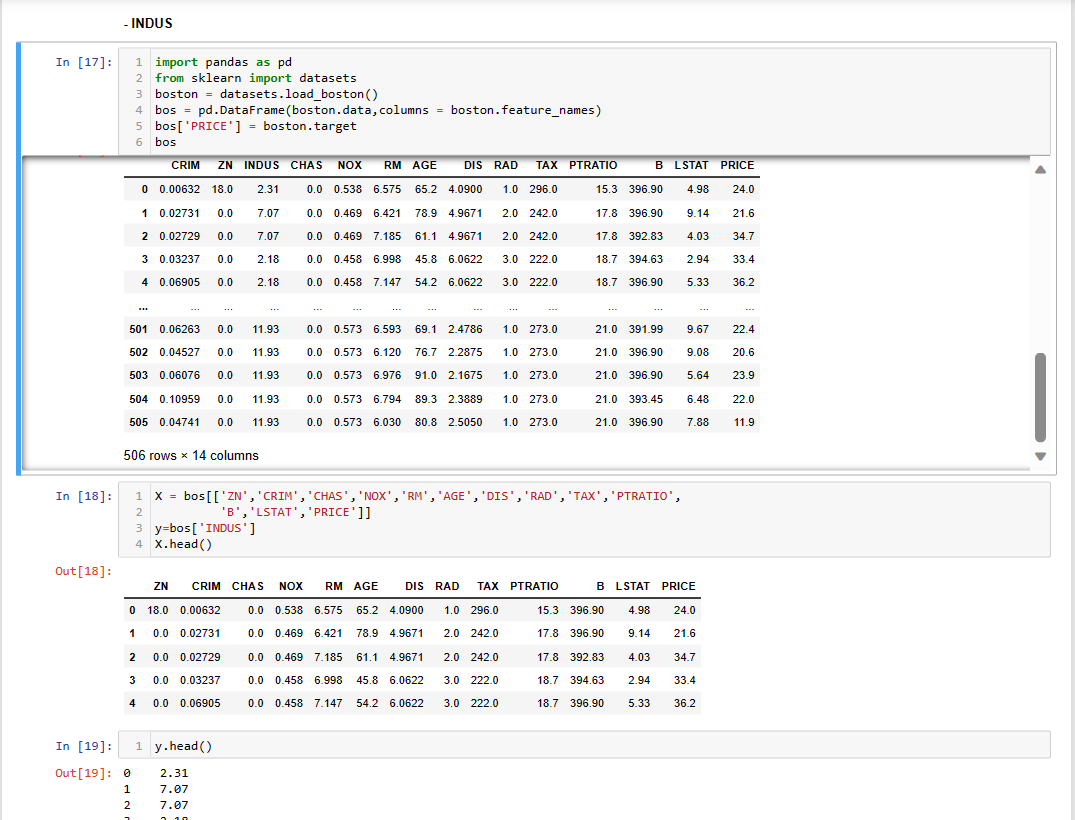
****

**- ZN**

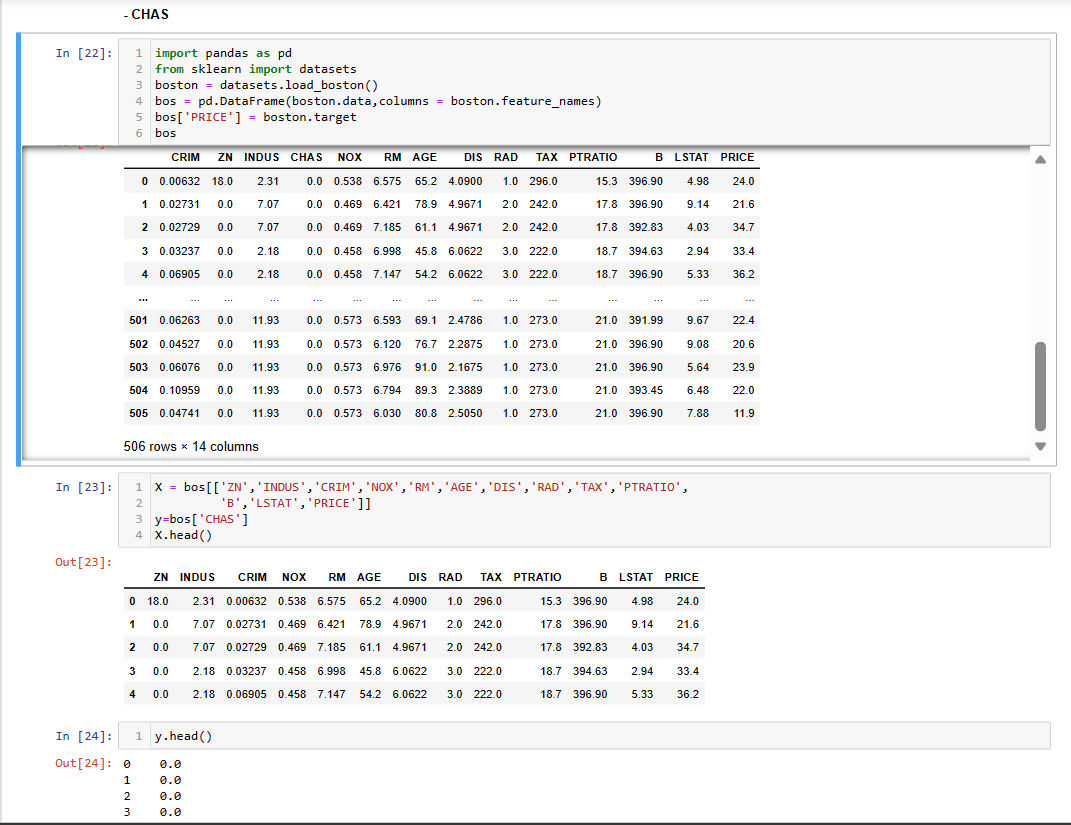
****

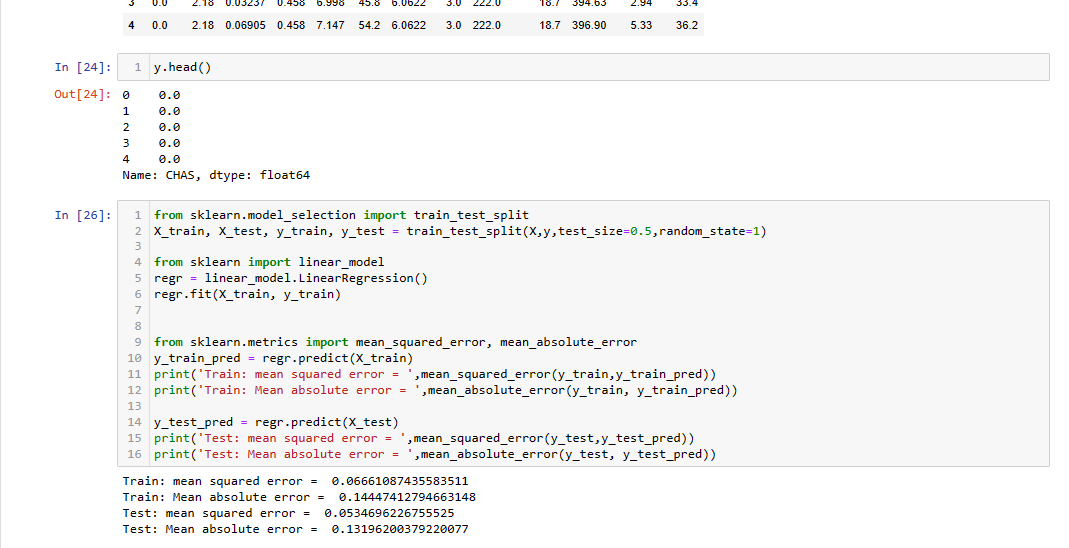
****

**- INDUS**

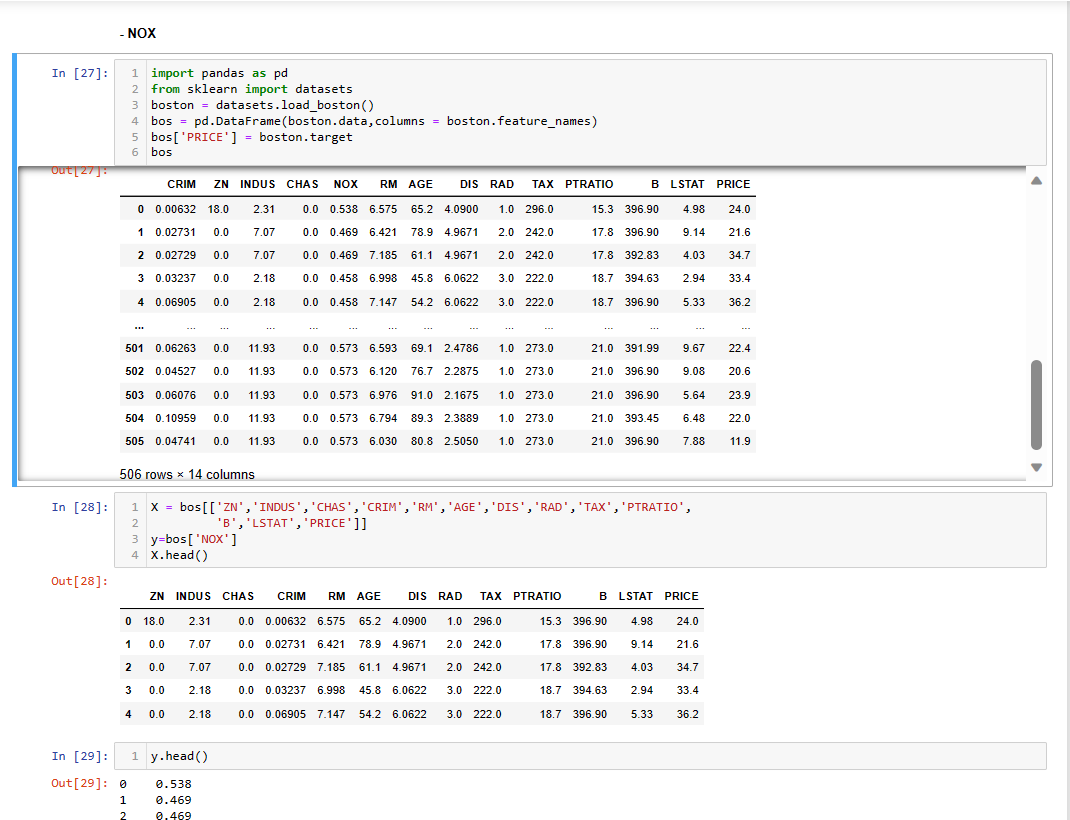
****

**- CHAS**

****

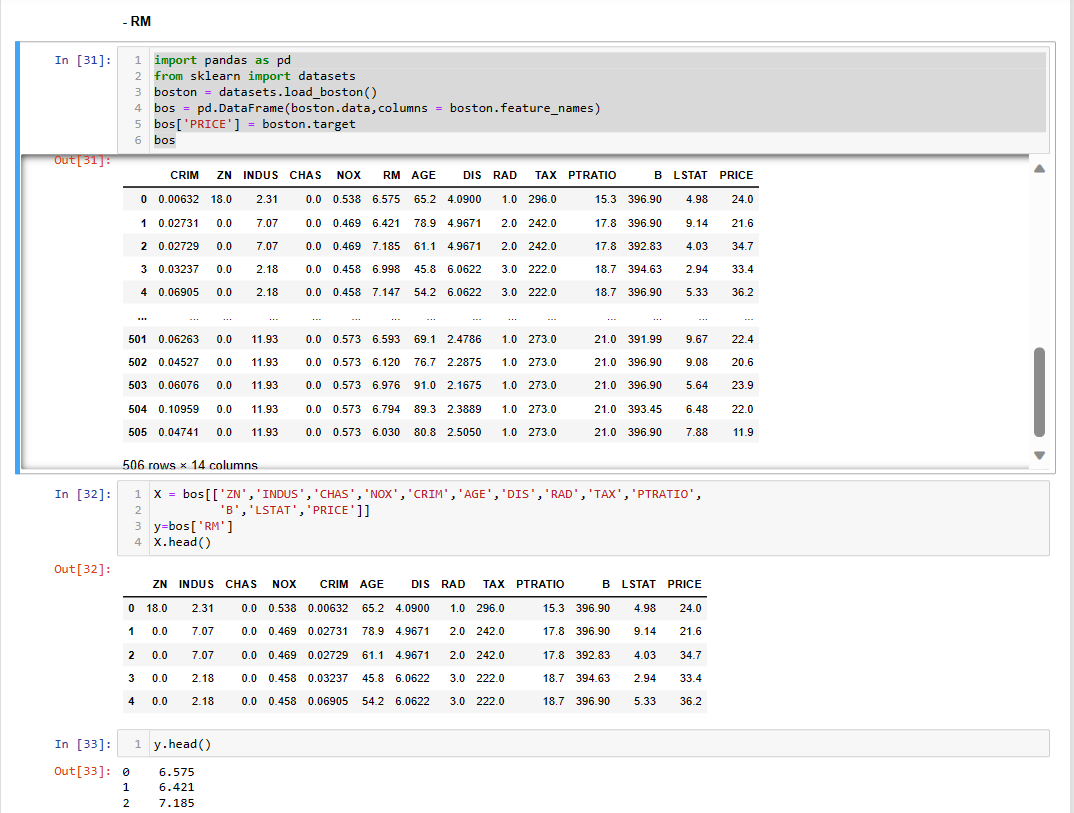
****

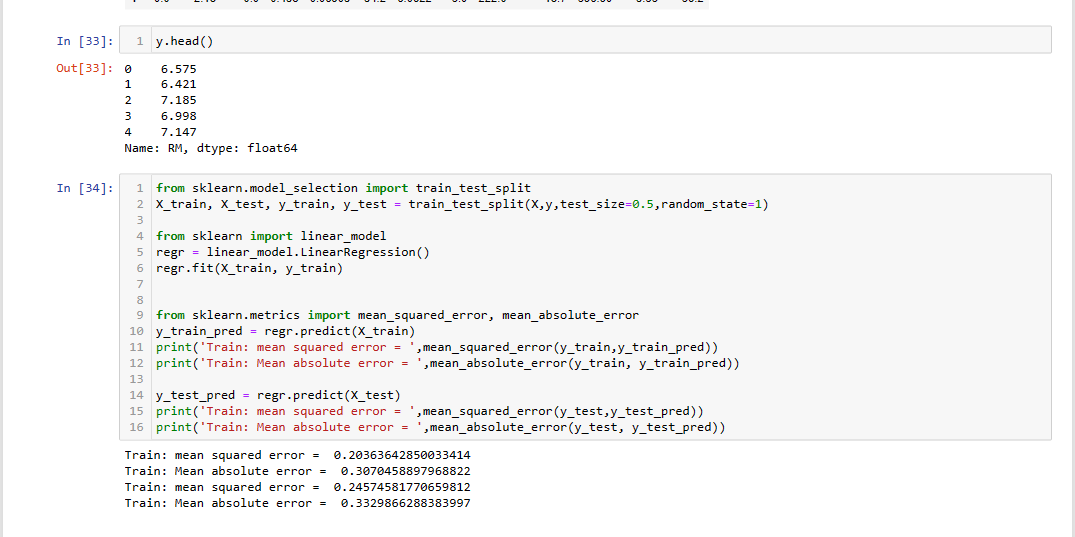
**- NOX**

****

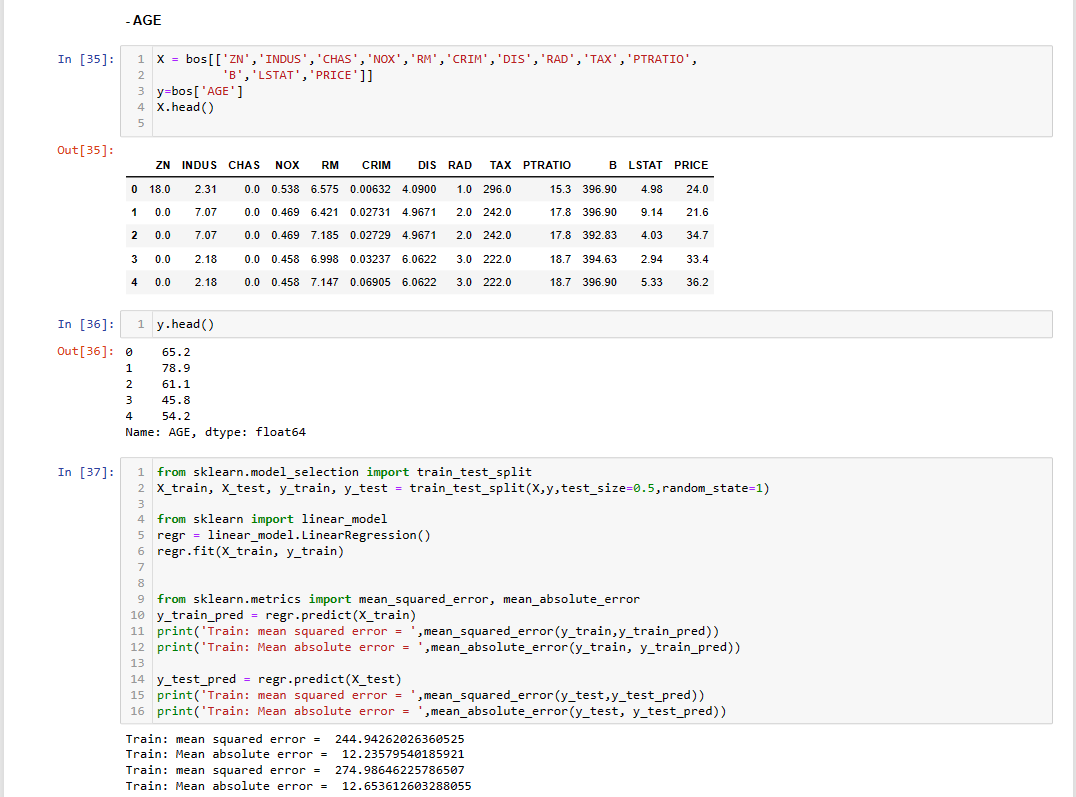
****

**- RM**

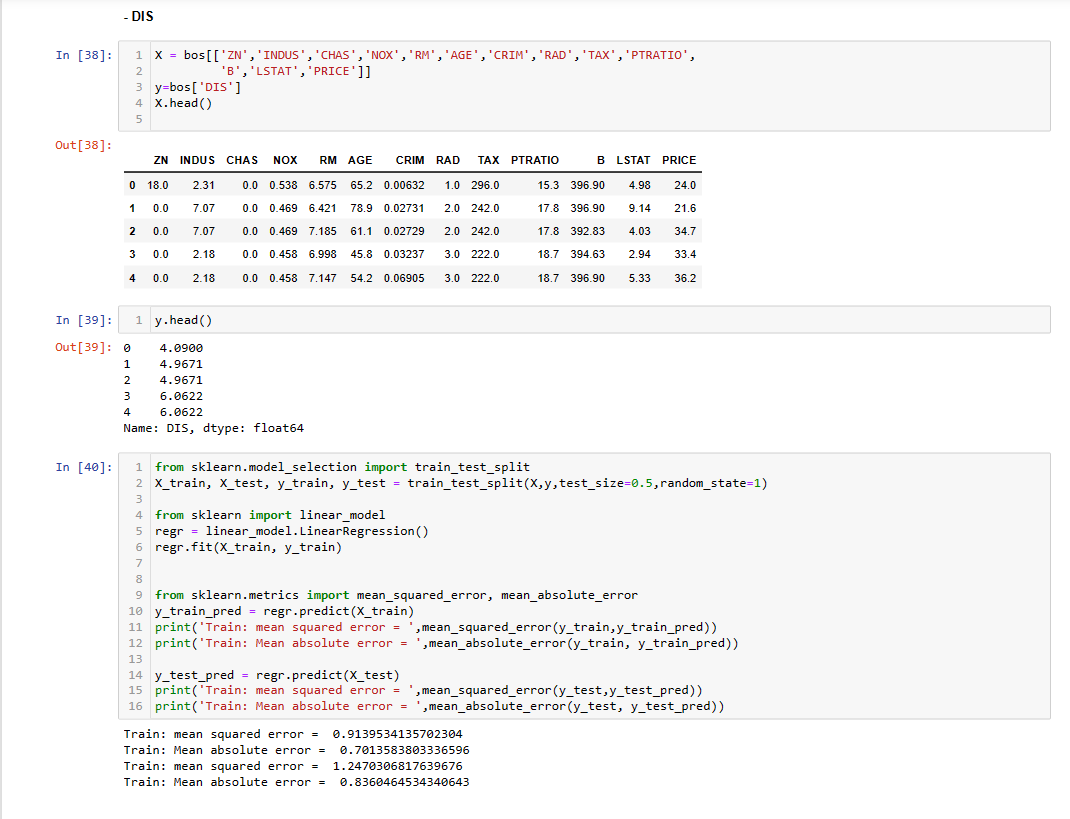
****

****

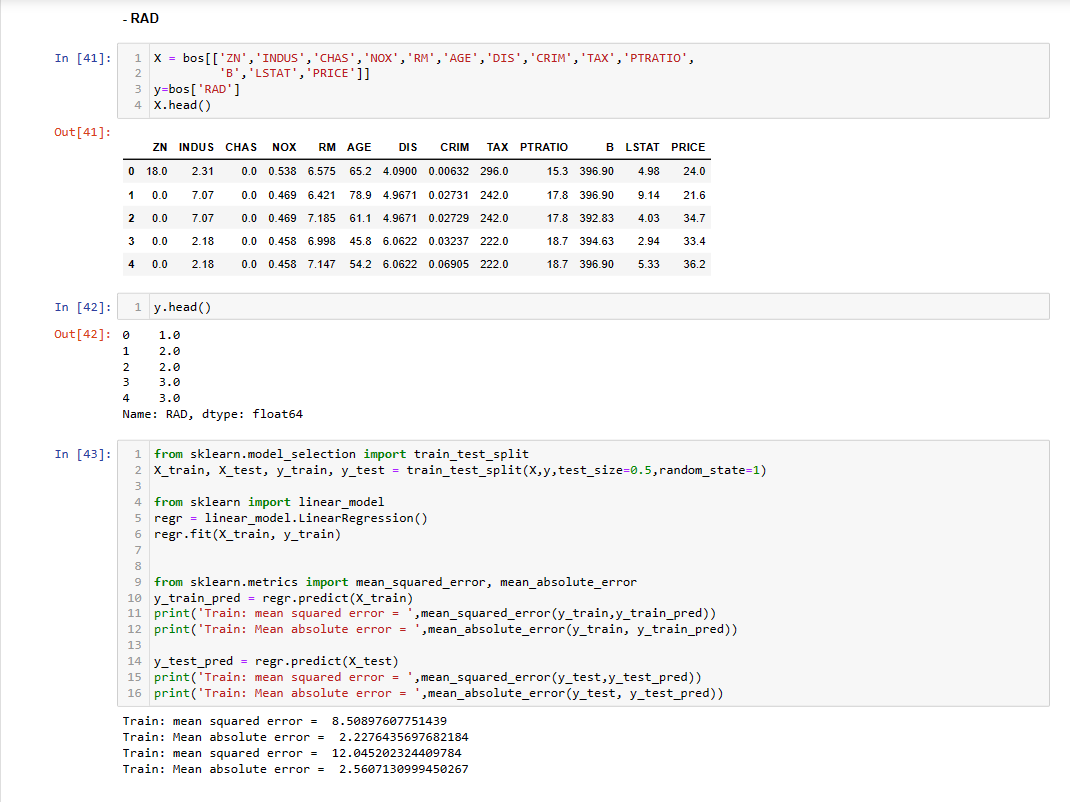
**- AGE**

****

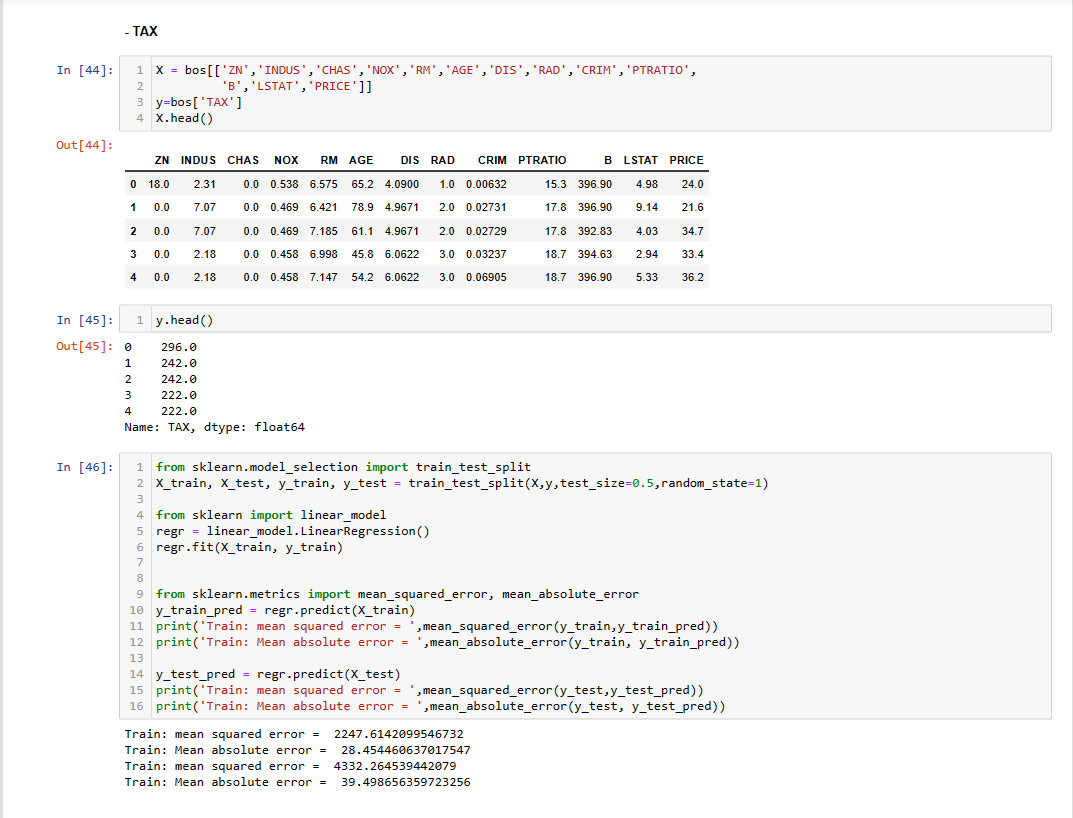
**- DIS**

****

**- RAD**

****

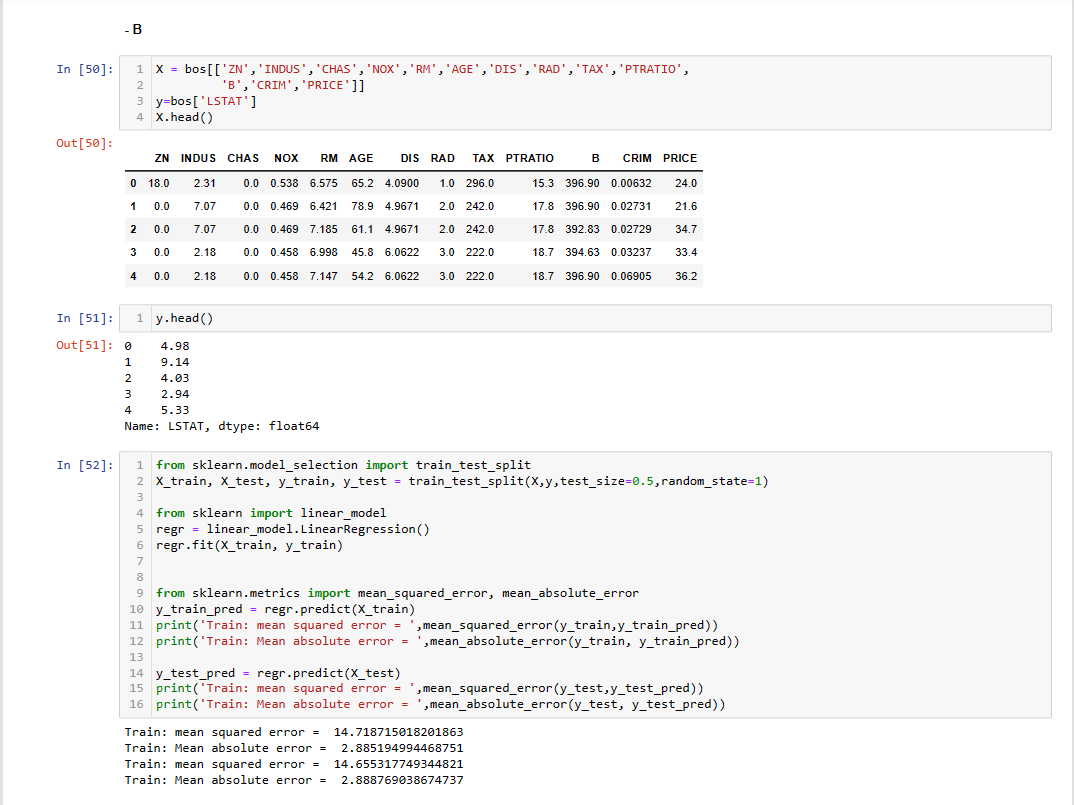
**- TAX**

****

**- PTRATIO**

****

**- B**

****

**- LSTAT**



1. Lập bảng tổng hợp kết quả so sánh kết quả huấn luyện trên tập huấn luyện và kiểm tra từ câu trên, nhận xét kết quả. (Tham khảo cách nhận xét kết quả trong bảng[[3]](#footnote-2).

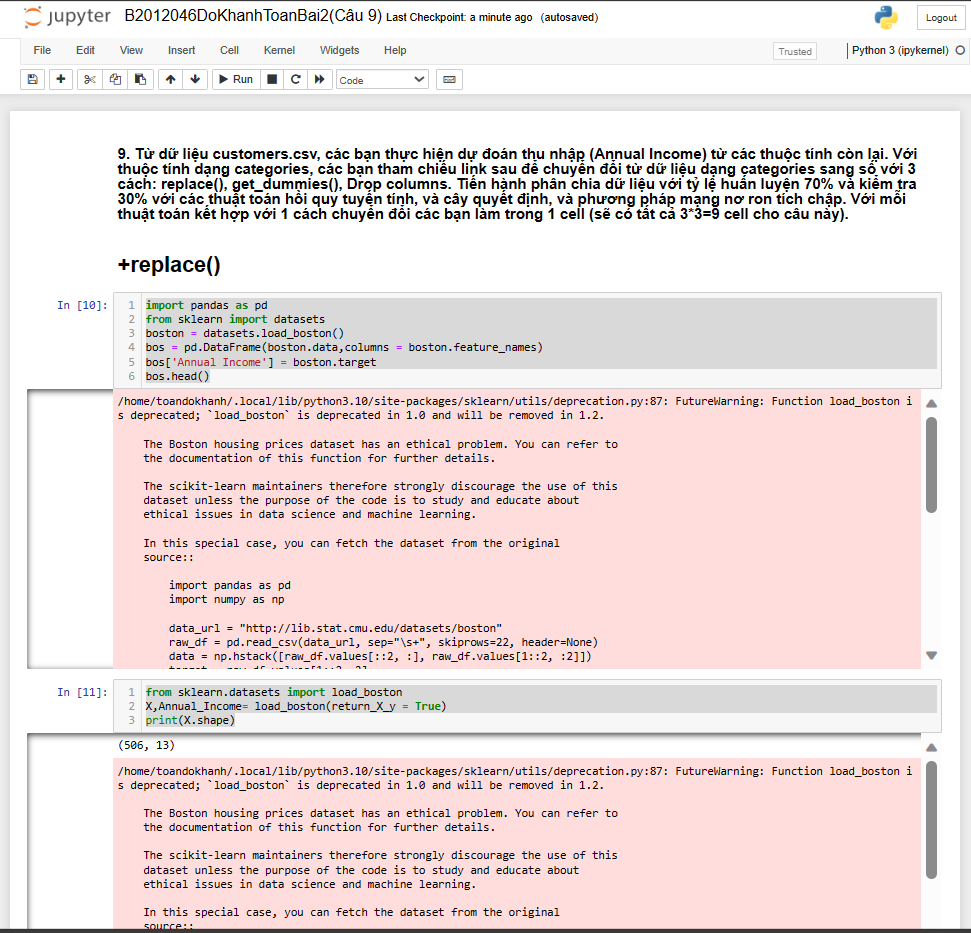
**Bài làm**

|  | **Train** | **Test** | **So sánh** |
| --- | --- | --- | --- |
| **CRIM** | mean squared error = 44.51688447794812  Mean absolute error = 3.3045123418530324 | Mean absolute error = 3.30451234185303247  Mean absolute error = 3.345230726891577 | Tập train cao hơn tập test  => kết qua không tốt lắm |
| **ZN** | Train: mean squared error = 173.82902750673807  Train: Mean absolute error = 9.269434503029249 | Test: mean squared error = 316.52905171427653  Test: Mean absolute error = 12.24312707620584 | Tập train cao hơn tập test quá nhiều  => kết qua không tốt |
| **INDUS** | Train: mean squared error = 8.937247842241243  Train: Mean absolute error = 2.058867697737134 | Test: mean squared error = 15.600546426015852  Test: Mean absolute error = 2.6156834690821738 | Tập train thấp hơn tập test  => kết qua tốt |
| **CHAS** | Test: mean squared error = 15.600546426015852  Test: Mean absolute error = 2.6156834690821738 | Test: mean squared error = 15.600546426015852  Test: Mean absolute error = 2.6156834690821738 | Tập train và tập test quá thấp  => kết qua không tốt |
| **NOX** | Train: mean squared error = 0.003245615984087241  Train: Mean absolute error = 0.043698869528455545 | Test: mean squared error = 0.002737804599458551  Test: Mean absolute error = 0.041652591735830045 | Tập train và tập test quá thấp  => kết qua không tốt |
| **RM** | Train: mean squared error = 0.20363642850033414  Train: Mean absolute error = 0.3070458897968822 | Train: mean squared error = 0.24574581770659812  Train: Mean absolute error = 0.3329866288383997 | Tập train và tập test quá thấp  => kết qua không tốt |
| **AGE** | Train: mean squared error = 244.94262026360525  Train: Mean absolute error = 12.23579540185921 | Train: mean squared error = 274.98646225786507  Train: Mean absolute error = 12.653612603288055 | Tập train thấp hơn tập test  => kết qua tốt |
| **DIS** | Train: mean squared error = 274.98646225786507  Train: Mean absolute error = 12.653612603288055 | Train: mean squared error = 1.2470306817639676  Train: Mean absolute error = 0.8360464534340643 | Tập train và tập test quá thấp  => kết quá không tốt |
| **RAD** | Train: mean squared error = 8.50897607751439  Train: Mean absolute error = 2.2276435697682184 | Train: mean squared error = 12.045202324409784  Train: Mean absolute error = 2.5607130999450267 | Tập train thấp hơn tập test  => kết qua tốt |
| **TAX** | Train: mean squared error = 2247.6142099546732  Train: Mean absolute error = 28.454460637017547 | Train: mean squared error = 2247.6142099546732  Train: Mean absolute error = 28.454460637017547 | Tập train thấp hơn tập test  => kết qua rất tốt |
| **PTRATIO** | Train: mean squared error = 2247.6142099546732  Train: Mean absolute error = 28.454460637017547 | Train: mean squared error = 2.9147392460273926  Train: Mean absolute error = 1.3710635618627975 | Tập train thấp hơn tập test  => kết qua tốt |
| **B** | Train: mean squared error = 14.718715018201863  Train: Mean absolute error = 2.885194994468751 | Train: mean squared error = 14.718715018201863  Train: Mean absolute error = 2.885194994468751 | Tập train thấp hơn tập test  => kết qua tốt |
| **LSTAT** | Train: mean squared error = 5706.107697222249  Train: Mean absolute error = 43.04735164345058 | Train: mean squared error = 5706.107697222249  Train: Mean absolute error = 43.04735164345058 | Tập train cao hơn tập test  => kết qua không tốt lắm |

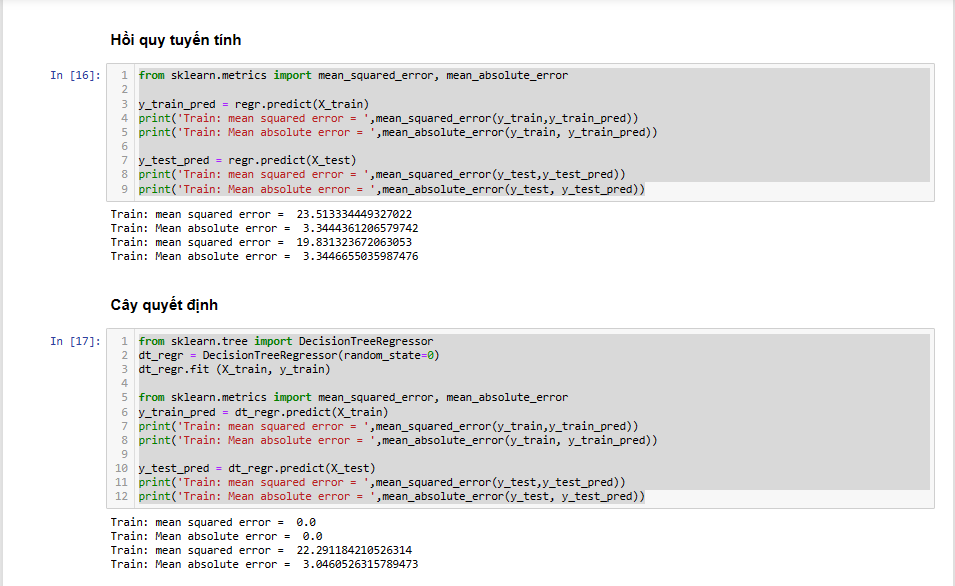
1. Từ dữ liệu customers.csv, các bạn thực hiện dự đoán thu nhập (Annual Income) từ các thuộc tính còn lại. Với thuộc tính dạng categories, các bạn tham chiếu link sau[[4]](#footnote-3) để chuyển đổi từ dữ liệu dạng categories sang số với 3 cách: replace(), get\_dummies(), Drop columns. Tiến hành phân chia dữ liệu với tỷ lệ huấn luyện 70% và kiểm tra 30% với các thuật toán hồi quy tuyến tính, và cây quyết định, và phương pháp mạng nơ ron tích chập. Với mỗi thuật toán kết hợp với 1 cách chuyển đổi các bạn làm trong 1 cell (sẽ có tất cả 3\*3=9 cell cho câu này).

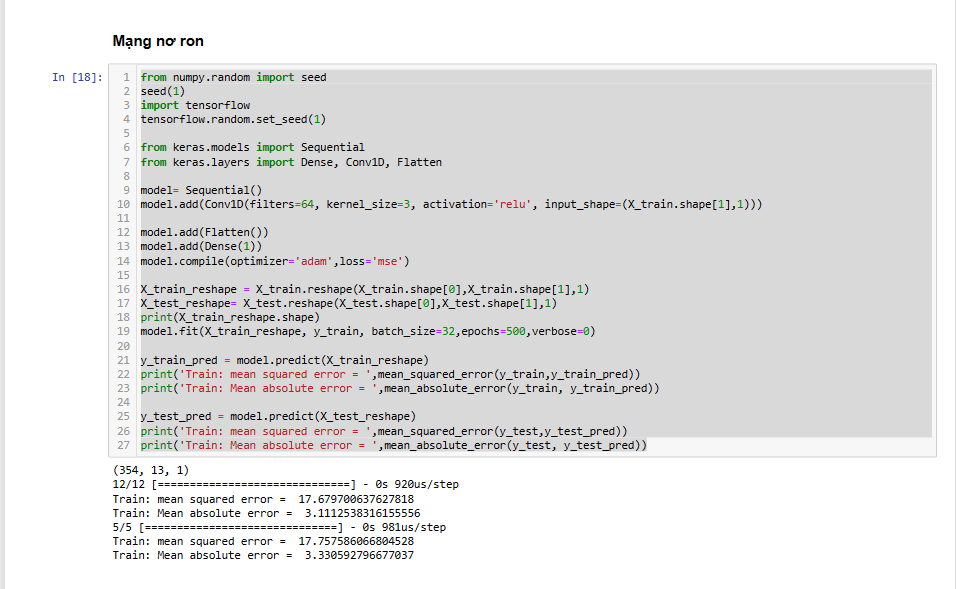
**Bài làm**

**REPLACE()**

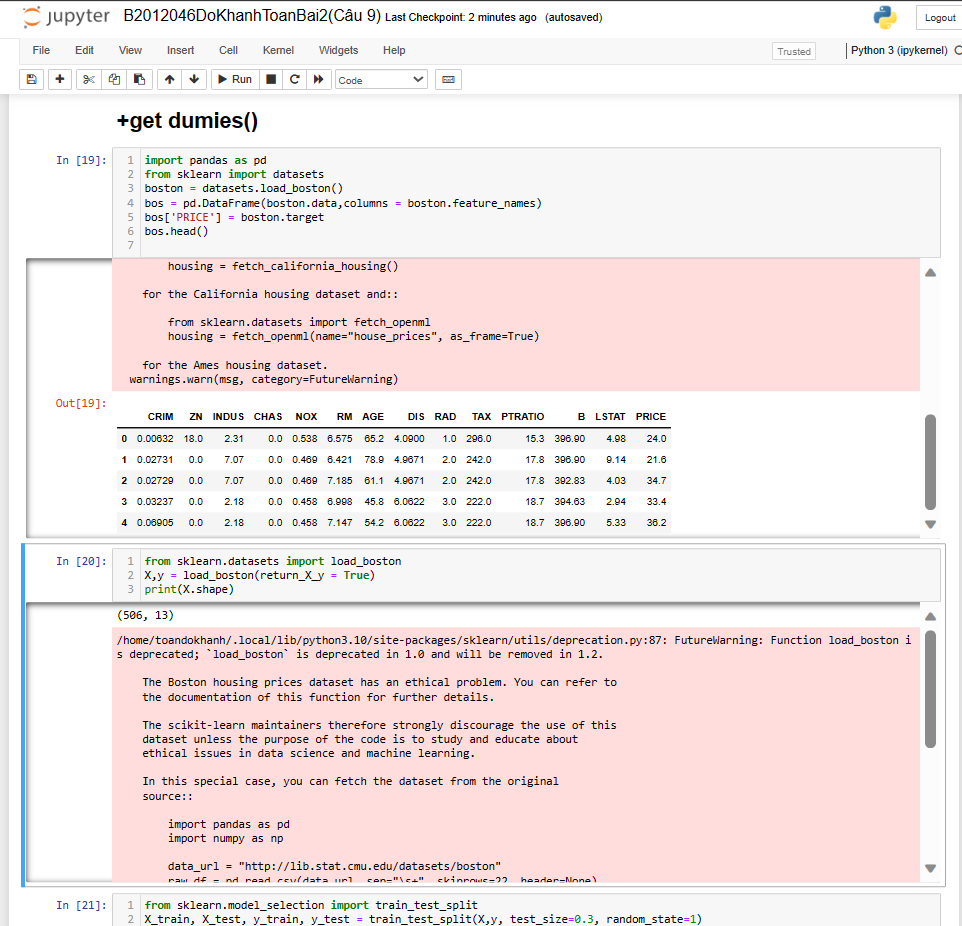


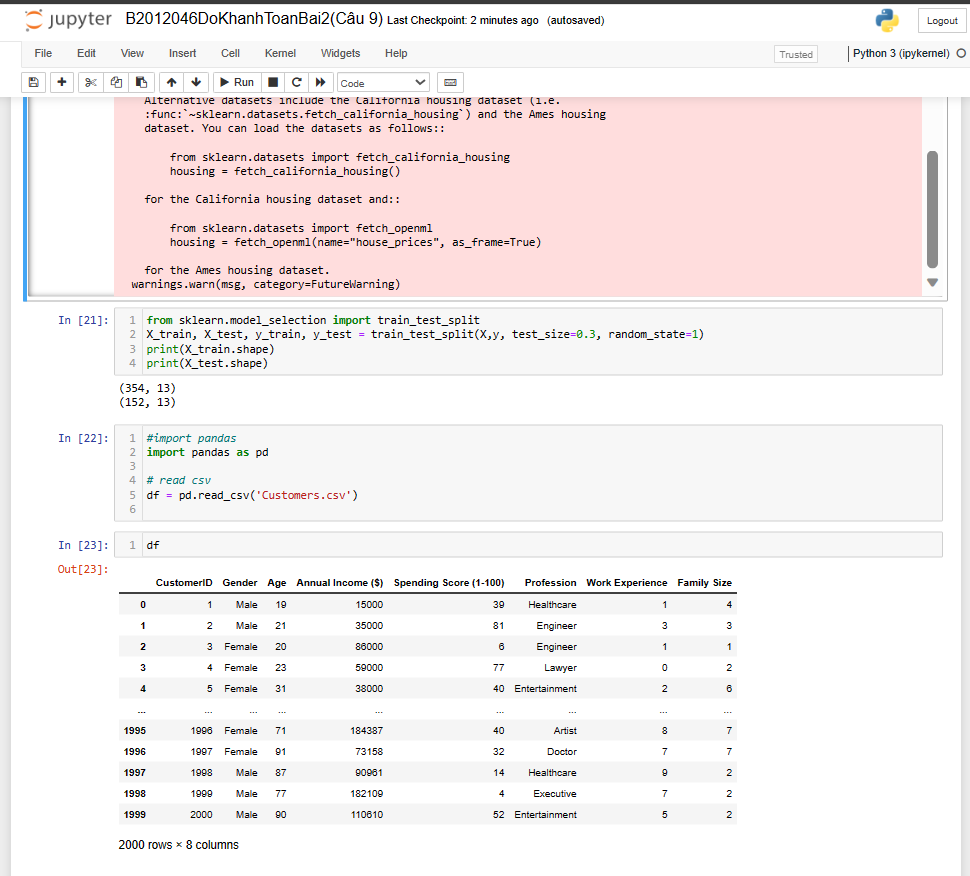




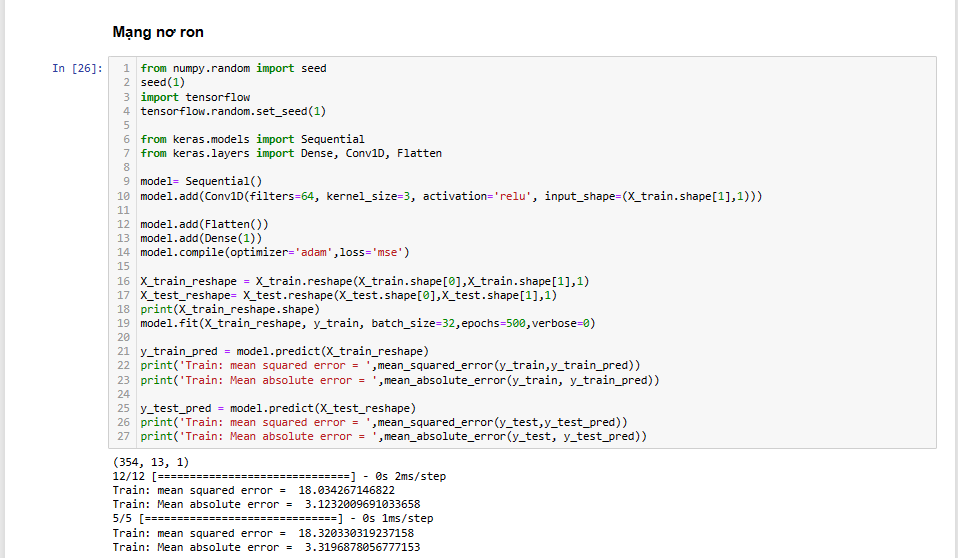


**GET DUMIES()**



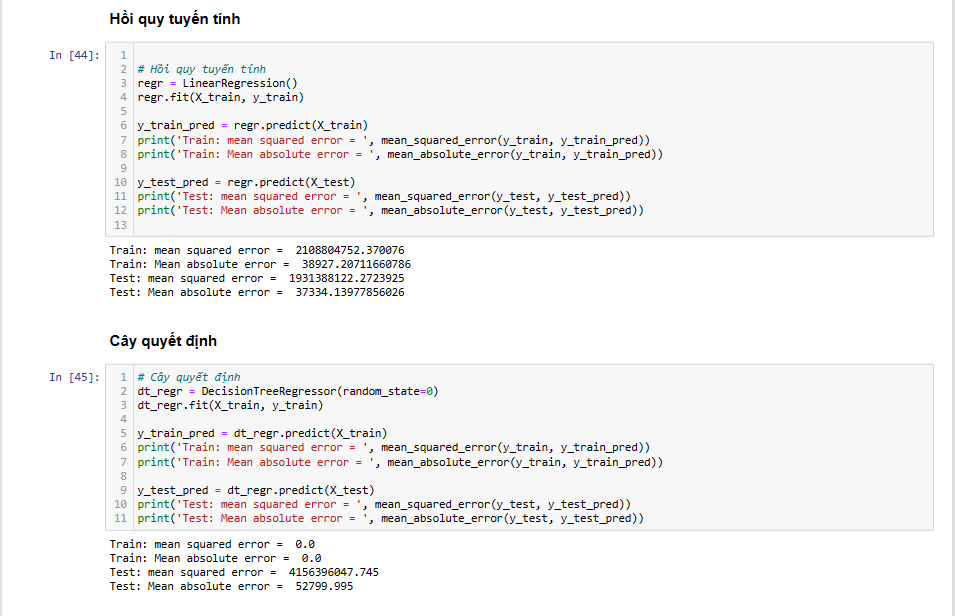


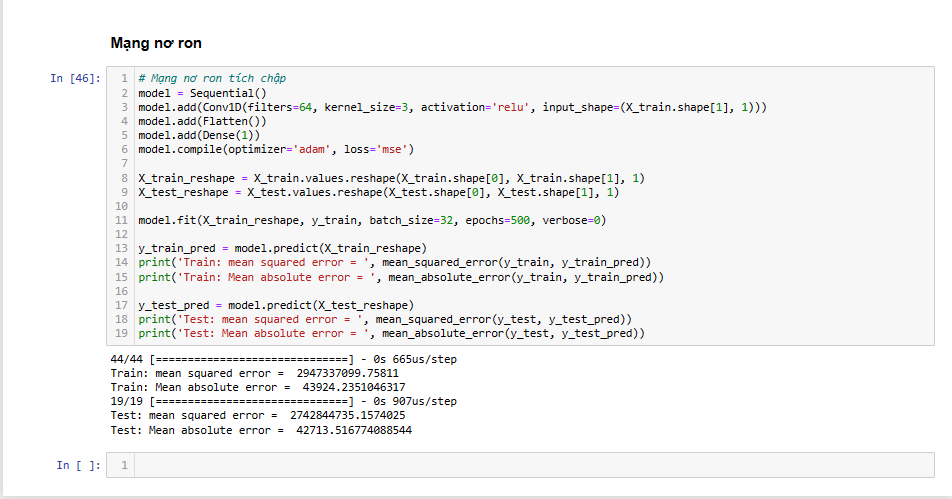




**DROP COLIMNS()**







1. Lập bảng tổng hợp kết quả so sánh kết quả huấn luyện trên tập huấn luyện và kiểm tra từ câu trên, nhận xét kết quả.

**Bài làm**

* Sau khi thay thế giá trị dự đoán (annual income) trong y với tỷ lệ tập huấn luyện là 70%, tập kiểm tra là 30% thì em thấy số dòng trong hai huấn luyện tăng lên và số cột thì không thay đổi. Và độ lỗi của tập huấn luyện thấp hơn độ lỗi trong tập kiểm tra, độ lỗi dự đoán của tập huấn luyện sau khi thay đổi đã tăng, độ lỗi thực của tập huấn luyện sau khi thay đổi đã tăng gần và độ lỗi dự đoán của tập kiểm tra sau khi thay đổi có xu hướng tăng

**Chú ý:**

* Các bạn nộp file word: Quy tắc đặt tên file: **<mssv>-<hoten>-<bai><stt\_bai thực hành>.docx** nộp lên Classroom (VD: **B123456-NguyenVanA-bai1.docx**), kèm với các file khác được yêu cầu như phần câu hỏi đã nêu. **Ngoại trừ file word trả lời câu hỏi, các file còn lại các bạn nén vào 1 file zip**. File zip đặt tên như file word.
* Mỗi câu các bạn trả lời bằng hình hoặc dạng text tùy vào yêu cầu của câu hỏi và **TRẢ LỜI THEO ĐÚNG THỨ TỰ CÂU HỎI**. Nếu câu nào không trả lời được các bạn cứ để số thứ tự câu hỏi và bỏ trống phần trả lời.
* Các câu trả lời có tham khảo trên Internet phải trích dẫn link/nguồn.
* *Vi phạm 1 trong các điều sau đây bài thực hành sẽ bị 0 điểm:*
  + Đặt tên KHÔNG ĐÚNG quy tắc được yêu cầu.
  + Bài không đủ các thành phần (word, code+data (nếu có),...) đã được yêu cầu.
  + Bài không thực hiện đúng yêu cầu “**Ngoại trừ file word trả lời câu hỏi, các file còn lại các bạn nén vào 1 file .zip”**
  + Bị phát hiện copy, sao chép từ các bạn khác
  + Phần trả lời không ghi rõ trả lời cho câu nào
  + Thứ tự câu trả lời không đúng thứ tự câu hỏi

1. https://www.sciencedirect.com/topics/computer-science/regression-task [↑](#footnote-ref-0)
2. https://scikit-learn.org/0.24/datasets/toy\_dataset.html#boston-dataset [↑](#footnote-ref-1)
3. https://www.idialy.com/2015/02/cach-nhan-xet.html) [↑](#footnote-ref-2)
4. https://www.geeksforgeeks.org/how-to-convert-categorical-variable-to-numeric-in-pandas/ [↑](#footnote-ref-3)