**Bài tập thực hành 03:**

Bài toán phân lớp

# **Mục tiêu cần đạt**:

Biết cách cài đặt/ứng dụng luồng xử lý dữ liệu chung cho bài toán phân lớp, và một số kiến thức khác:

* Phân chia kiểu cross-validation
* Các độ đo bao gồm Confusion matrix
* Đánh giá/nhận xét kết quả học trên các tập dữ liệu
* Hiểu số filter trong kiến trúc CNN

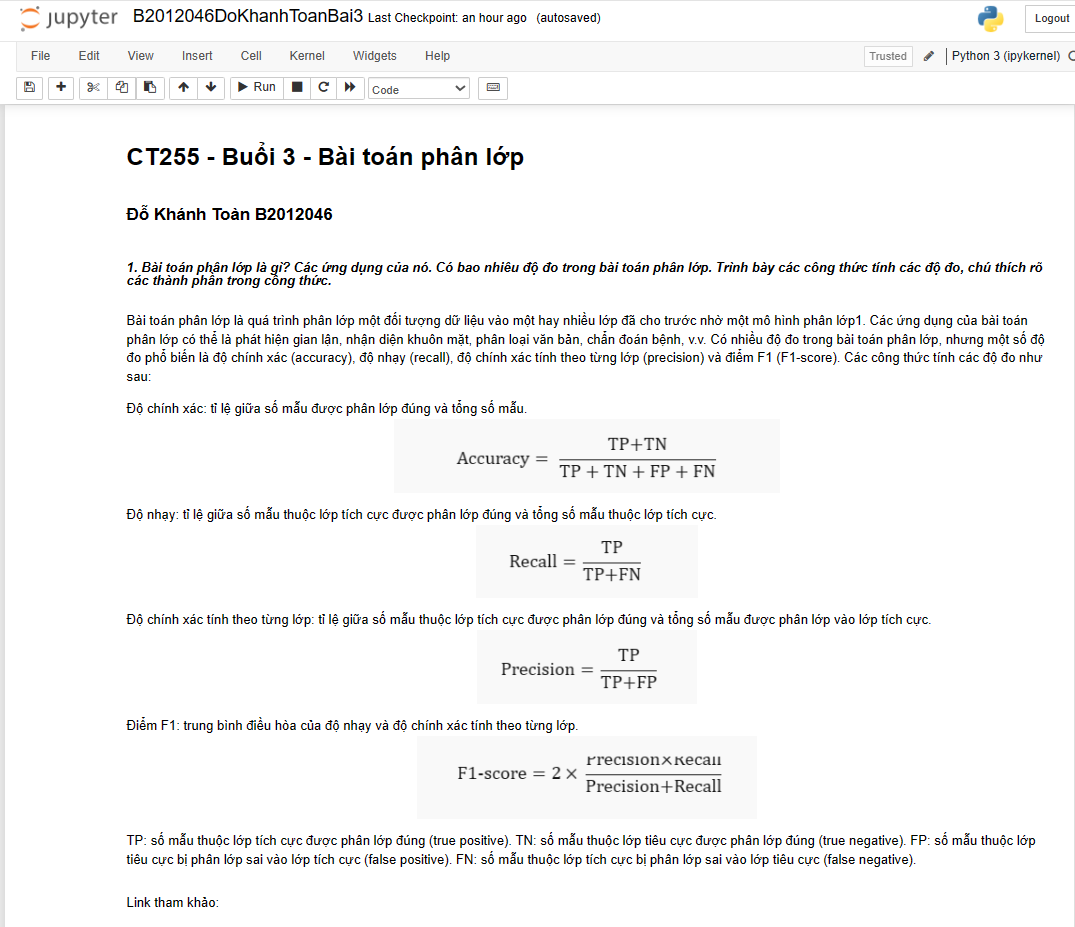
Các bạn thực hiện theo hướng dẫn trong giáo trình từ **trang 77 đến 84**, và trả lời các câu hỏi yêu cầu thực hành, có thể tham khảo thêm các link từ kaggle, github hoặc các nguồn khác để hỗ trợ trả lời câu hỏi (nhớ trích dẫn).

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

***Bài làm***



Link tham khảo:

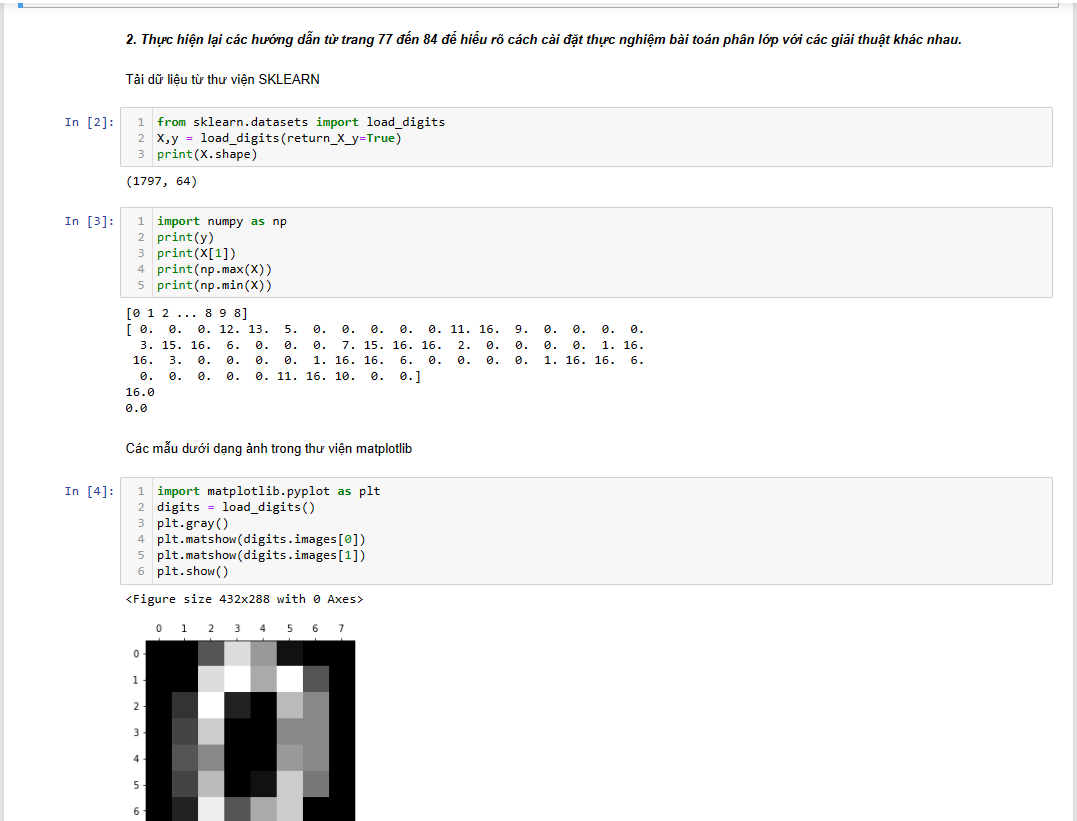
https://www.iostream.vn/article/bai-toan-phan-lop-trong-machine-learning-classification-in-machine-learning-515OIh

https://viblo.asia/p/thuat-toan-phan-lop-naive-bayes-924lJWPm5PM

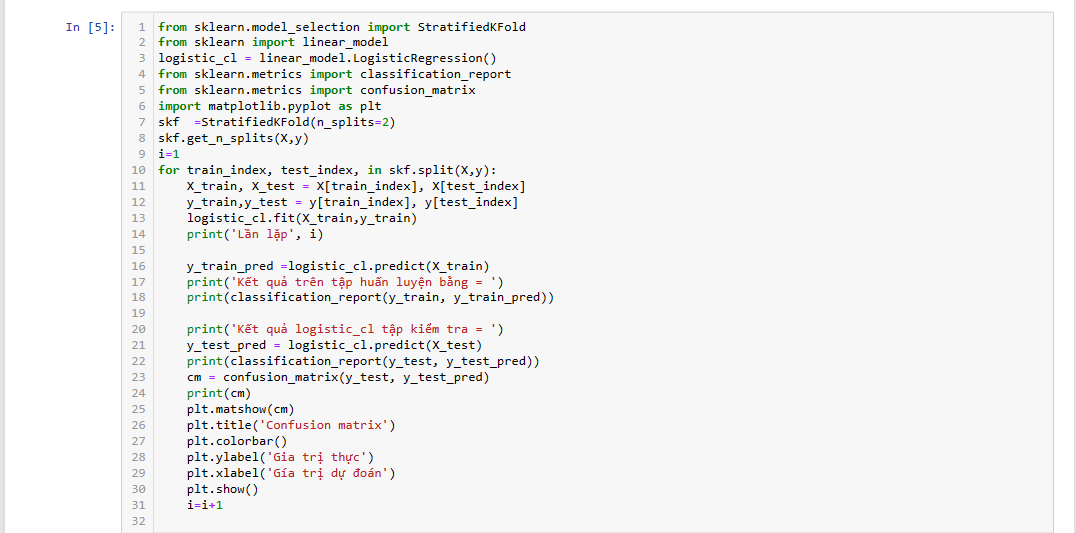
<https://vndoc.com/tuyen-tap-cac-bai-bai-toan-ve-phan-so-lop-4-5-123286>

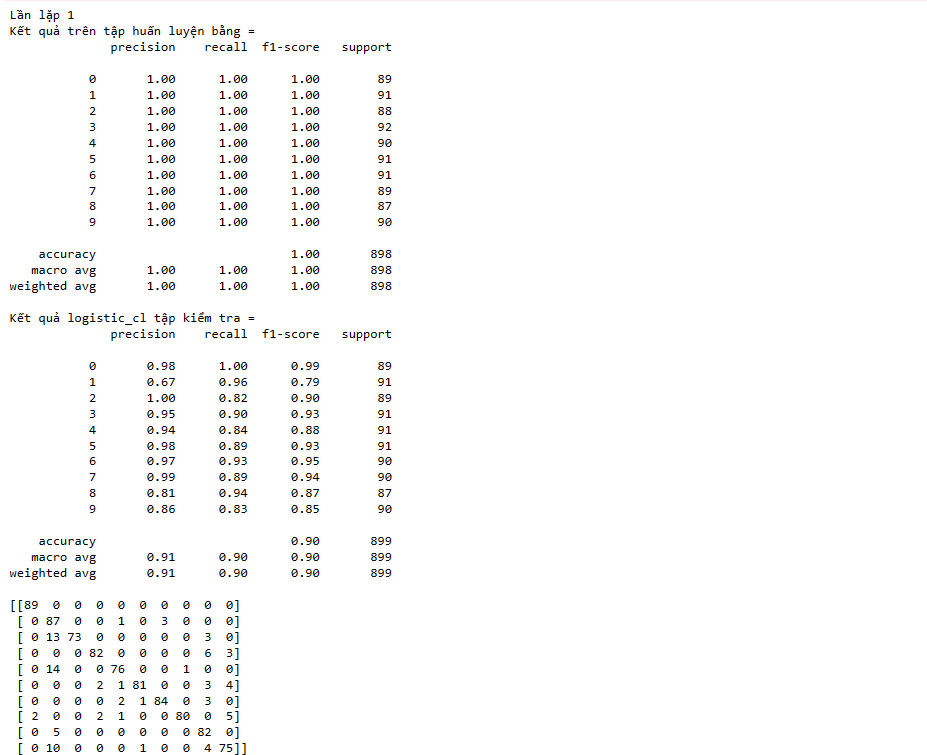
1. Thực hiện lại các hướng dẫn từ trang 77 đến 84 để hiểu rõ cách cài đặt thực nghiệm bài toán phân lớp với các giải thuật khác nhau.

***Bài làm***

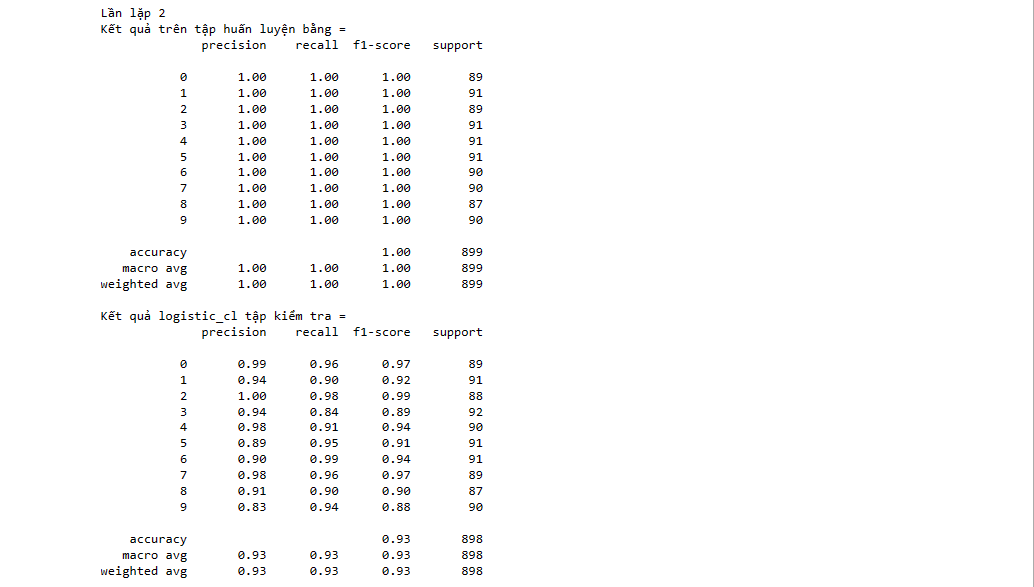


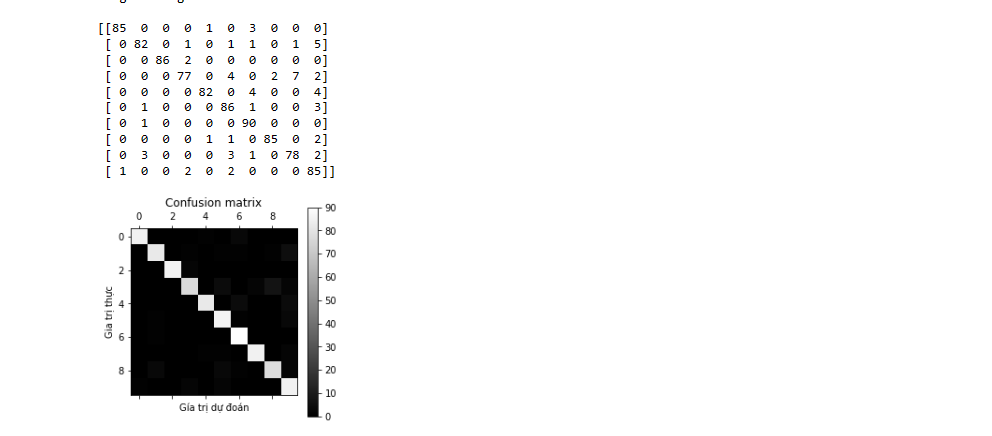


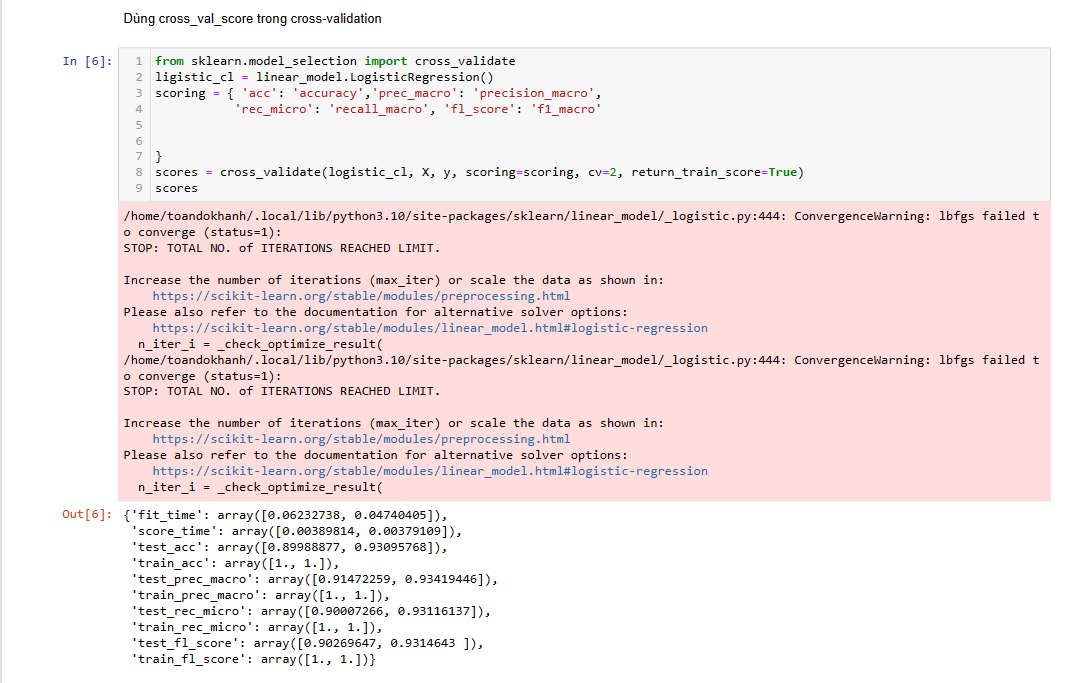


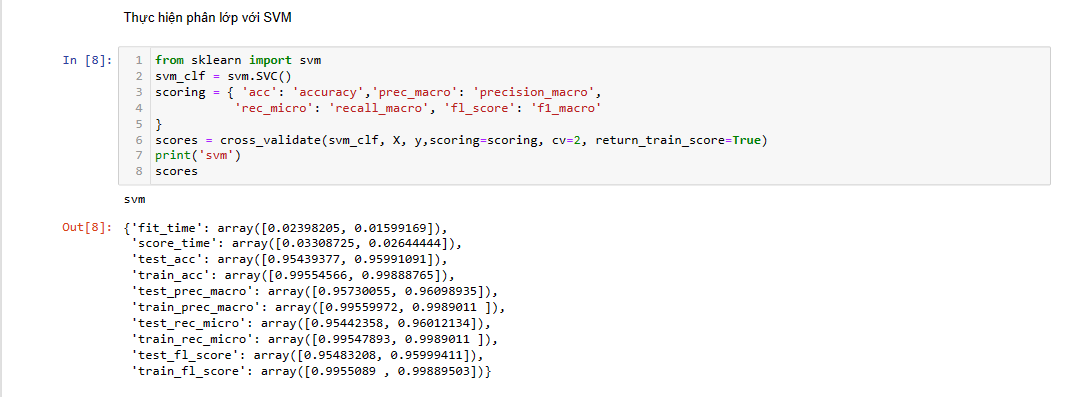


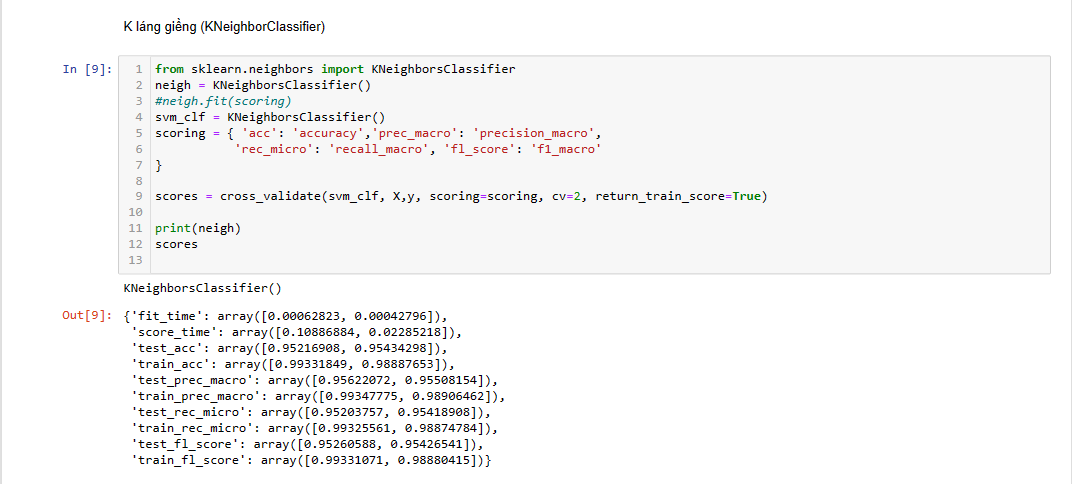


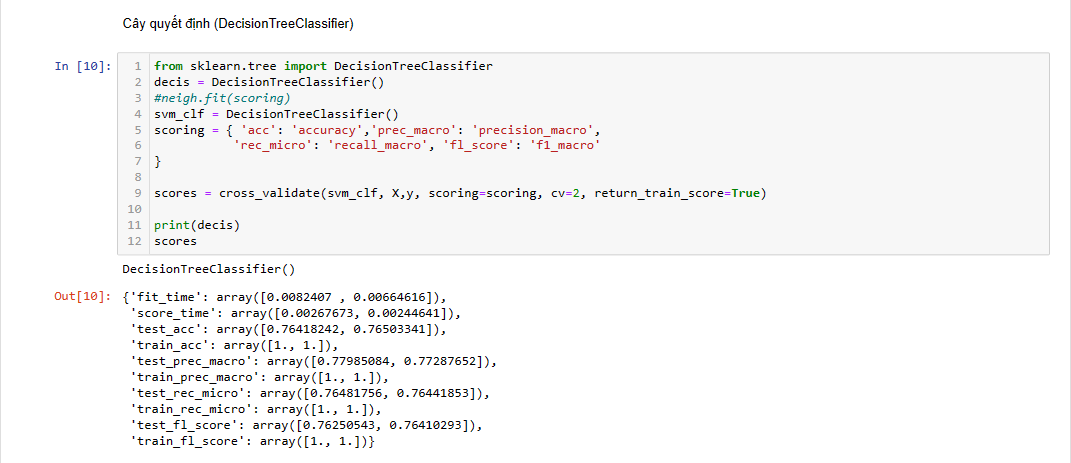


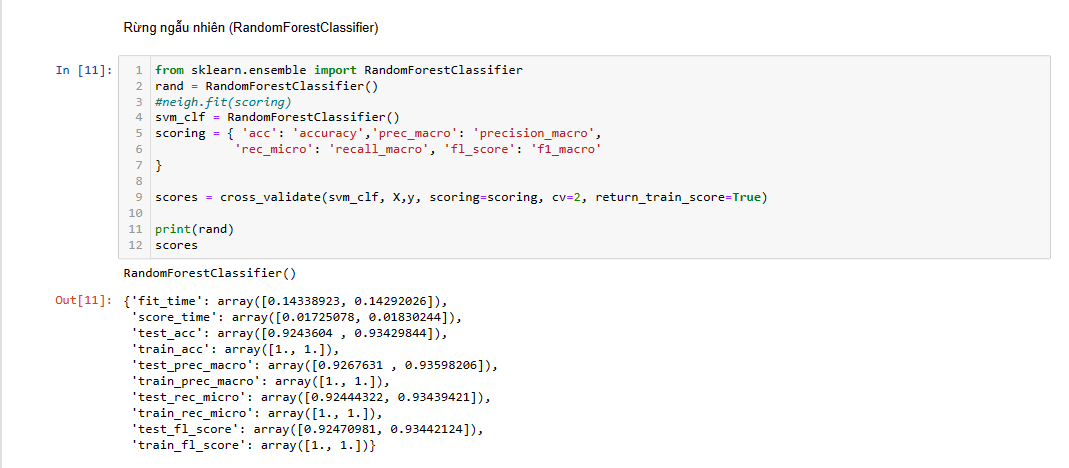


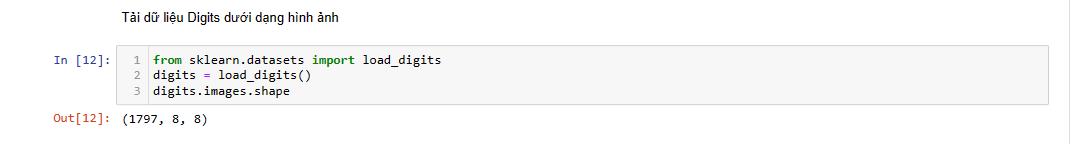




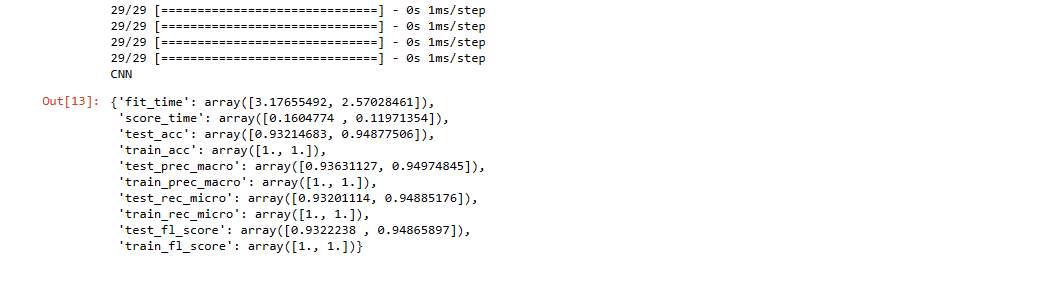






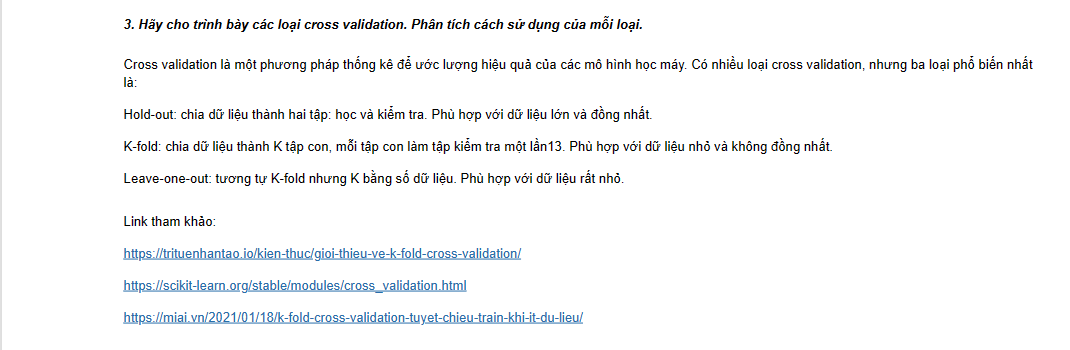






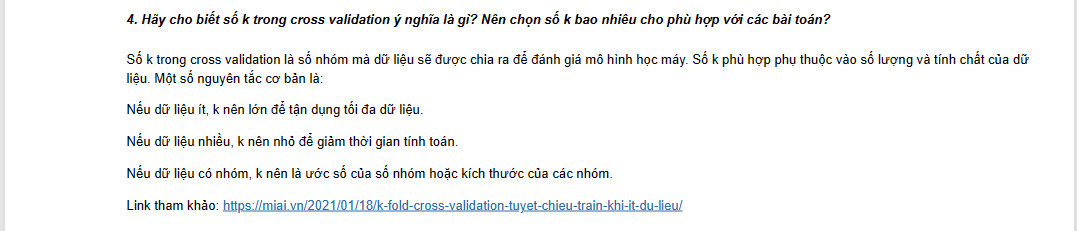
1. Hãy cho trình bày các loại cross validation. Phân tích cách sử dụng của mỗi loại.

***Bài làm***



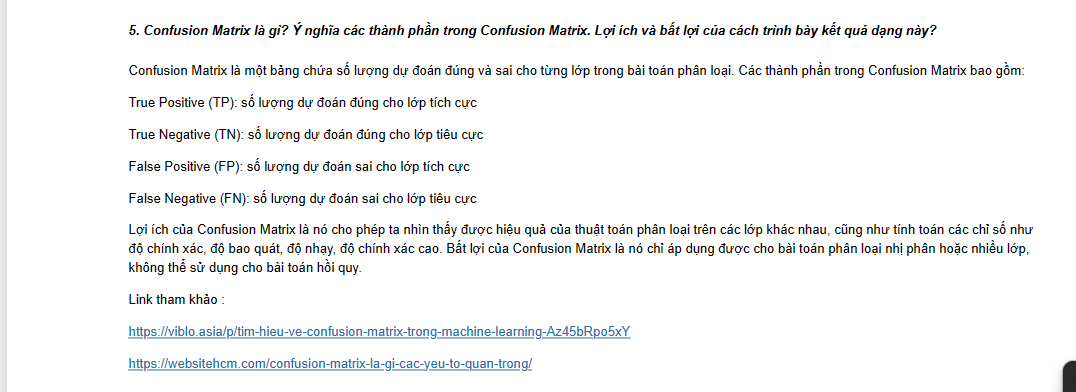
1. Hãy cho biết số k trong cross validation ý nghĩa là gì? Nên chọn số k bao nhiêu cho phù hợp với các bài toán?

***Bài làm***

******

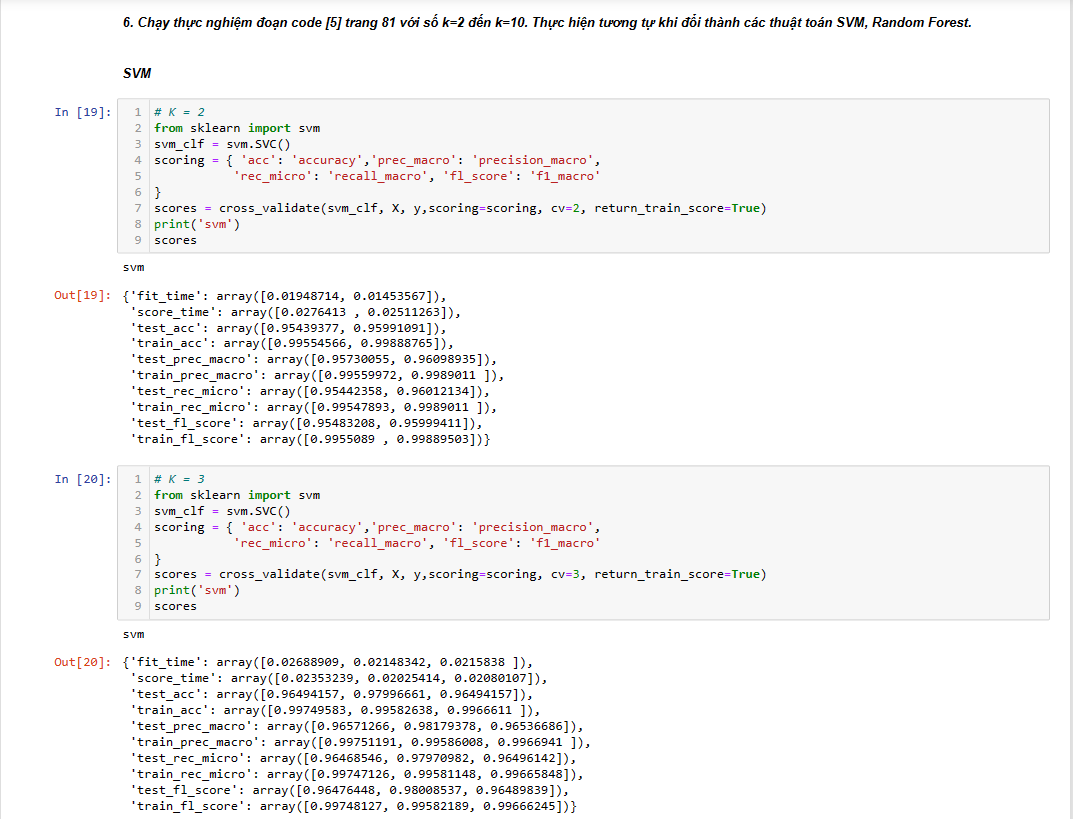
1. Confusion Matrix là gì? Ý nghĩa các thành phần trong Confusion Matrix. Lợi ích và bất lợi của cách trình bày kết quả dạng này?

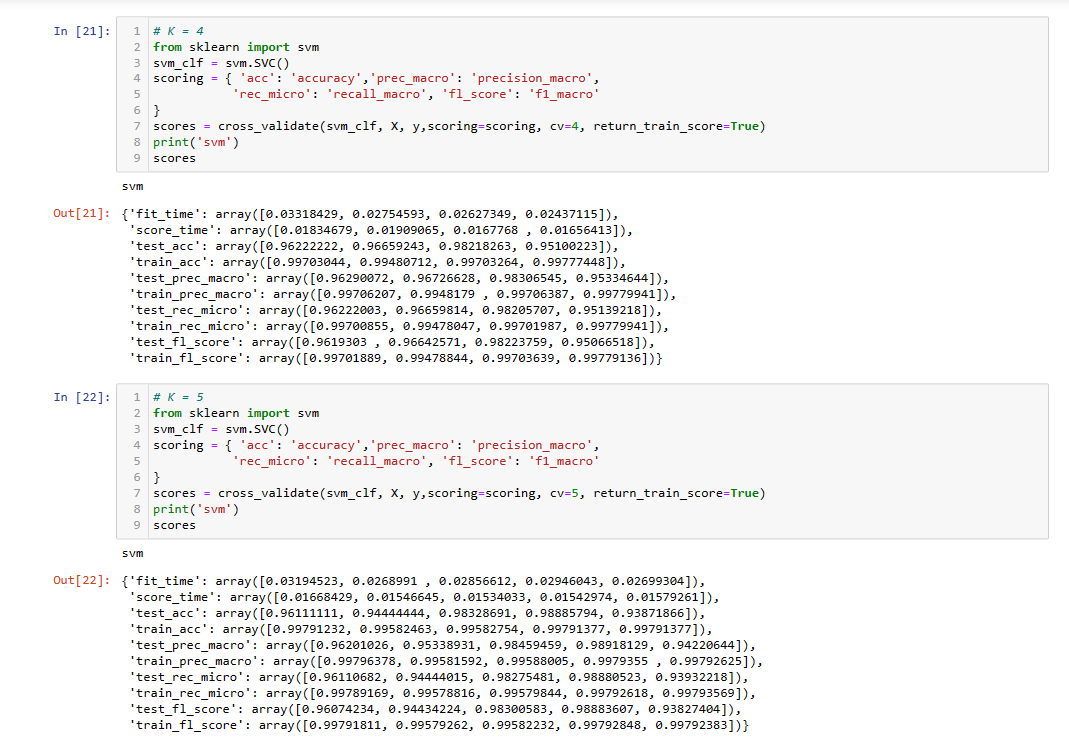
***Bài làm***

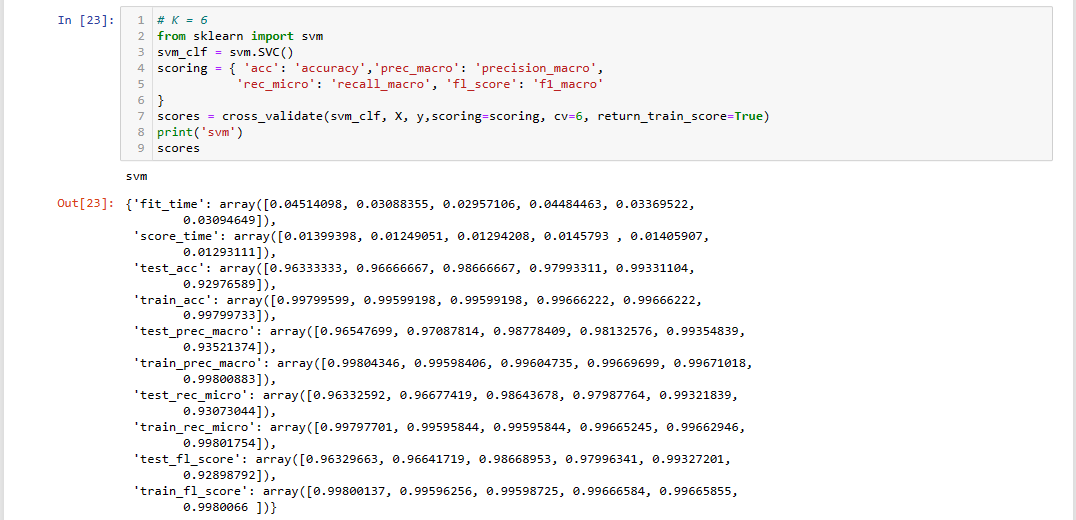
******

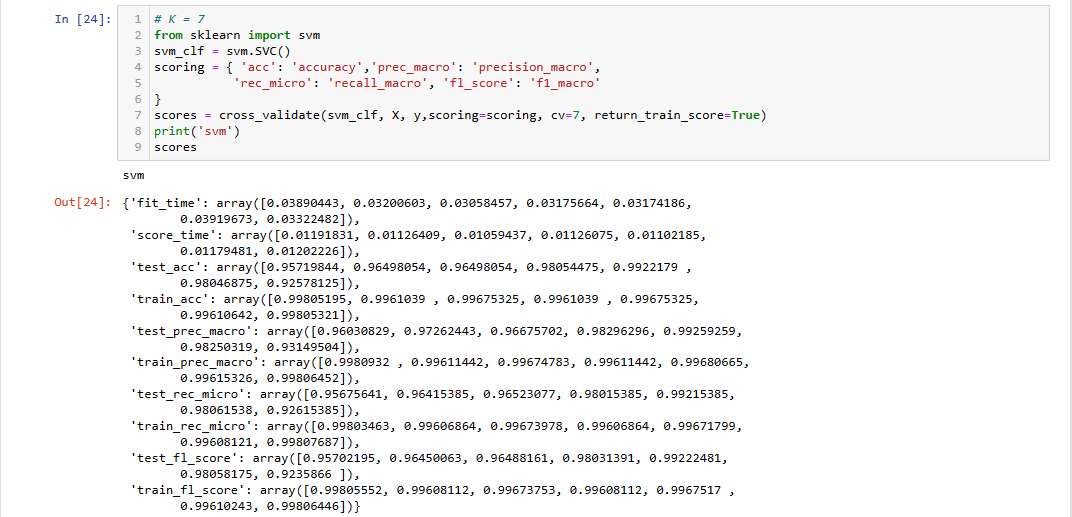
1. Chạy thực nghiệm đoạn code [5] trang 81 với số k=2 đến k=10. Thực hiện tương tự khi đổi thành các thuật toán SVM, Random Forest.

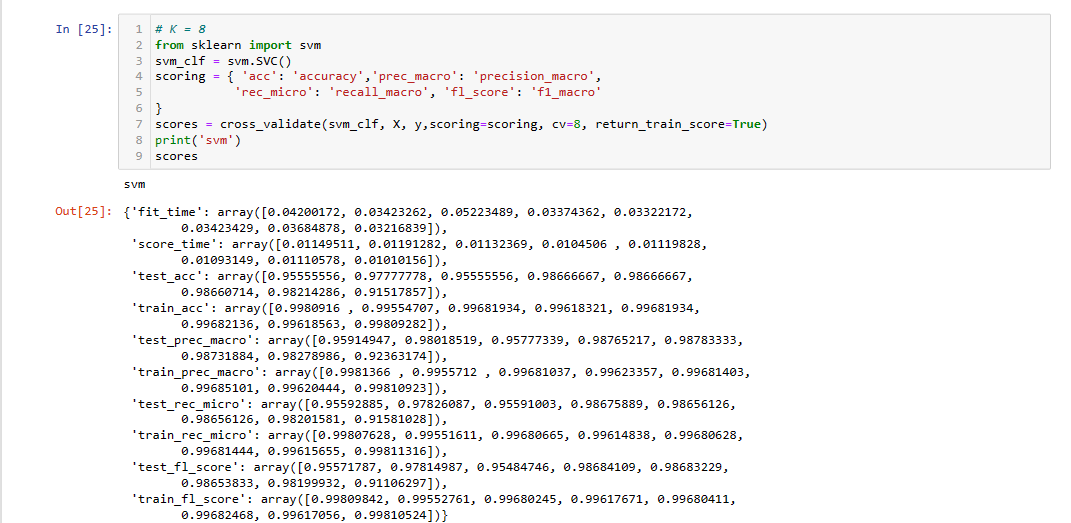
***Bài làm***

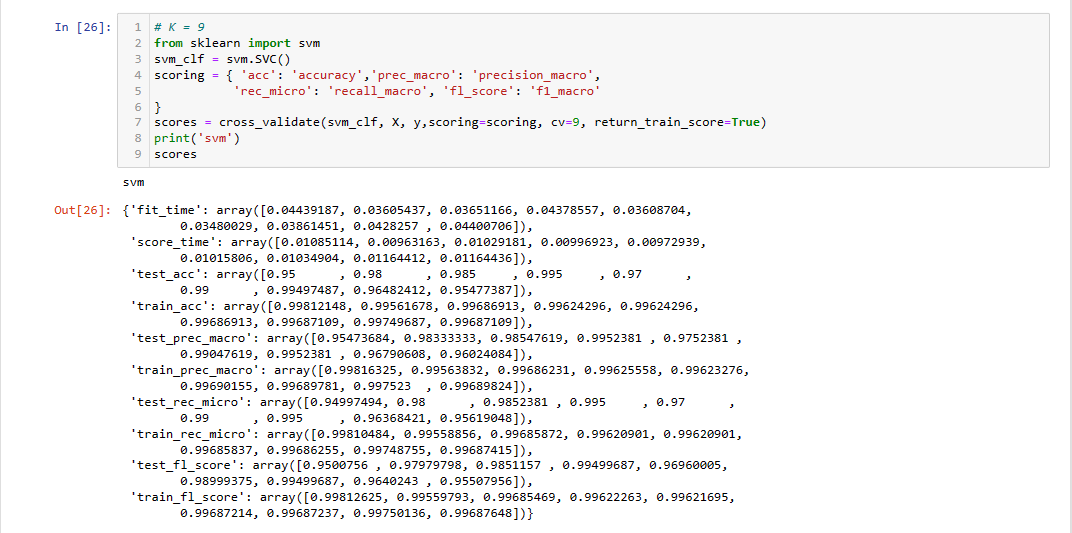
******

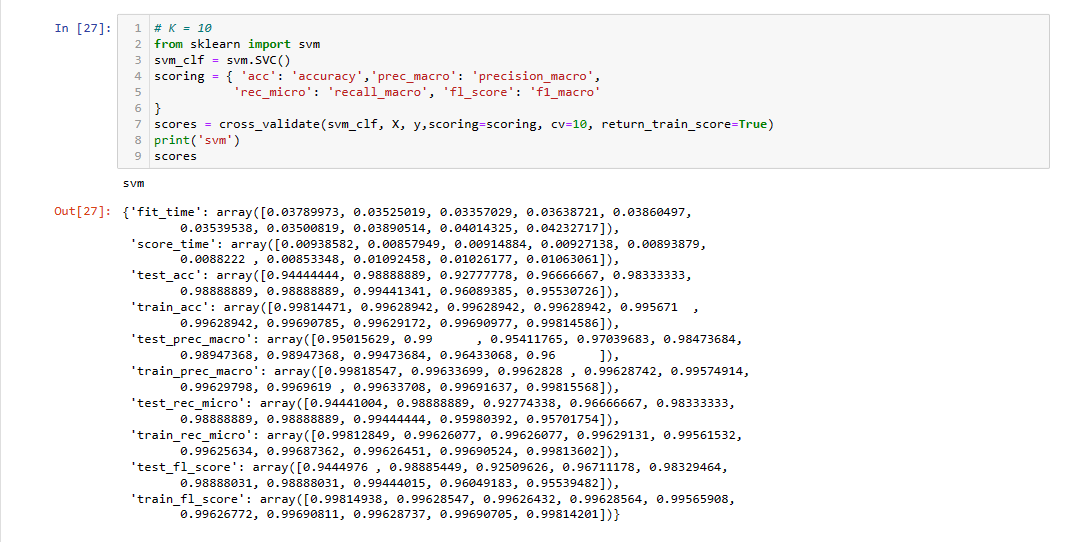
******

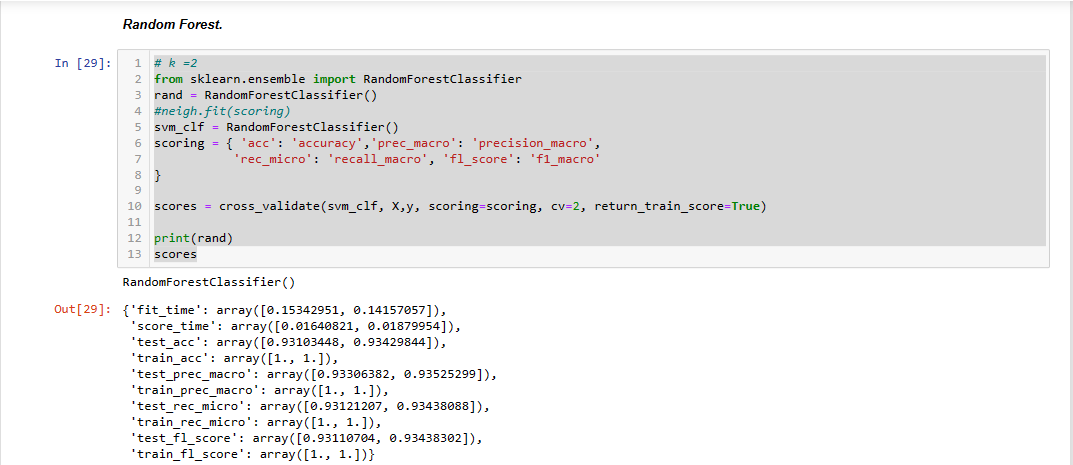
******

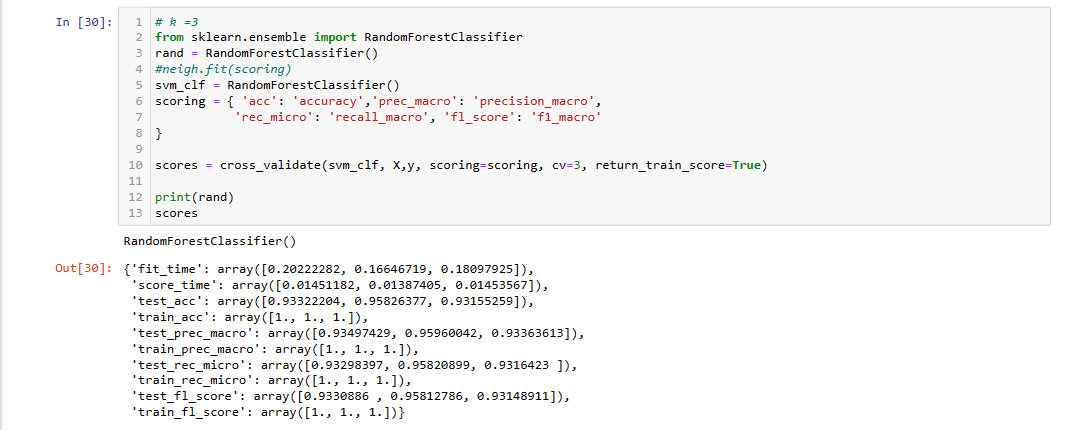
******

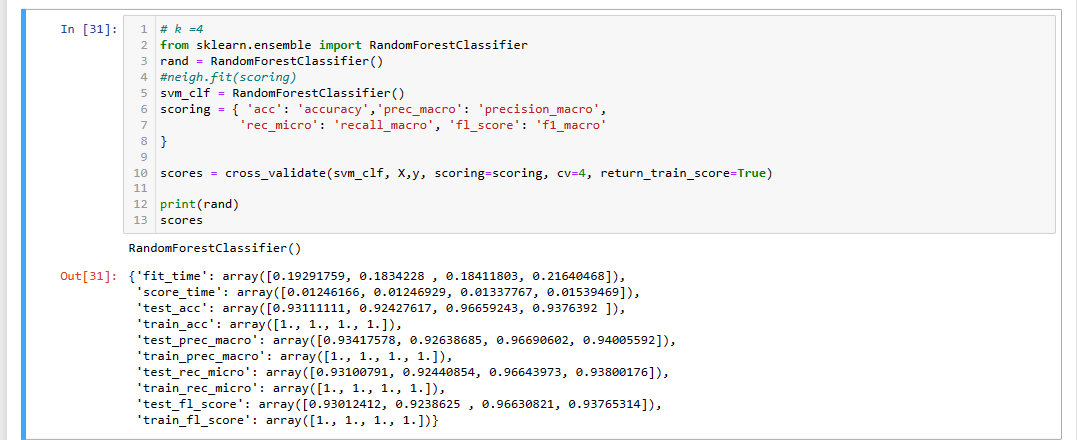
******

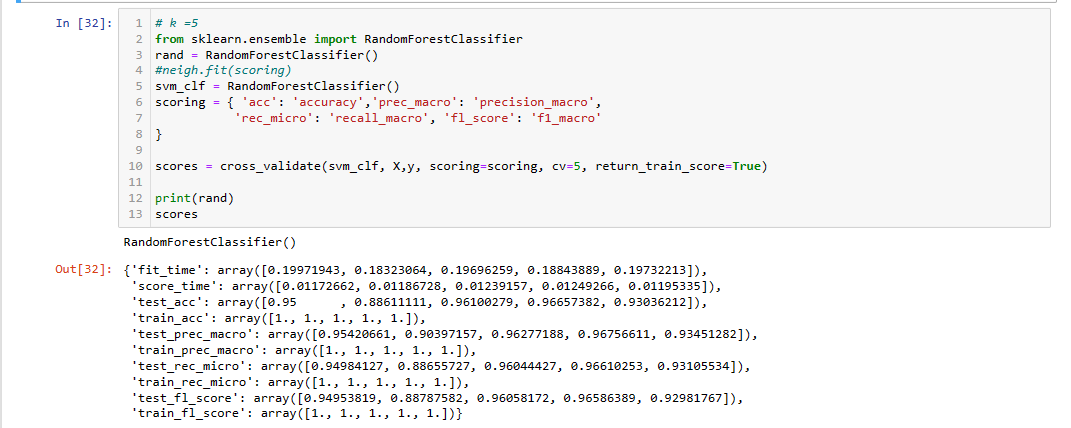
******

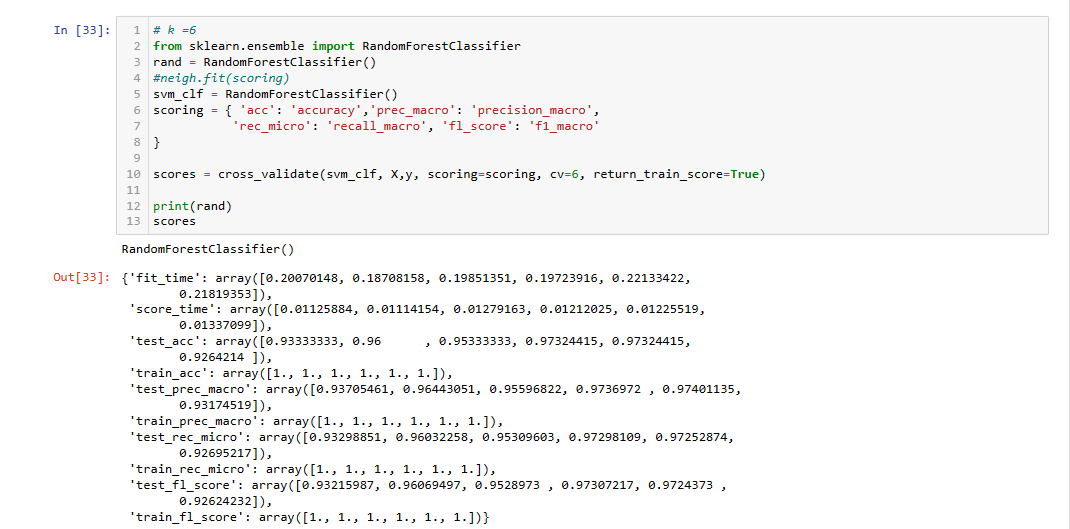
******

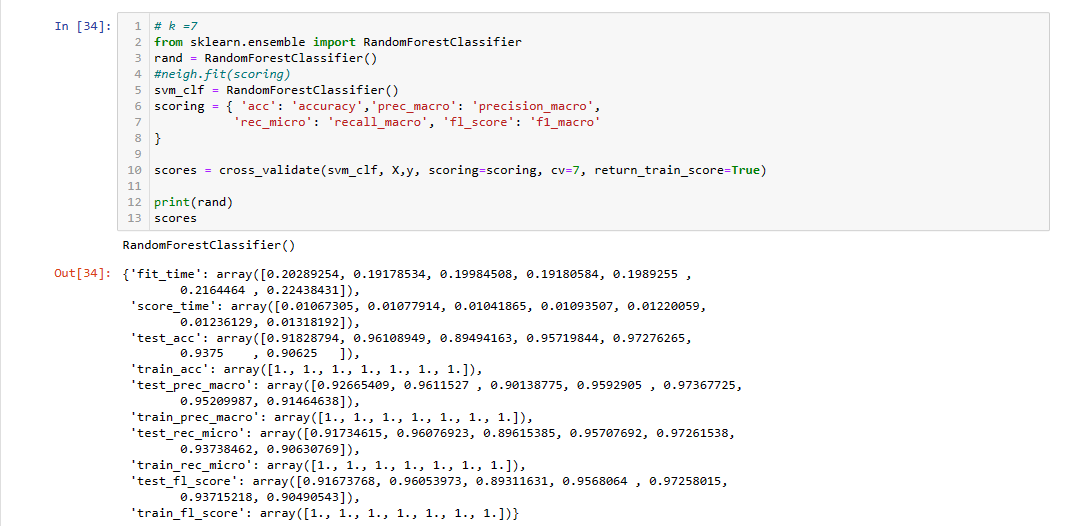
******

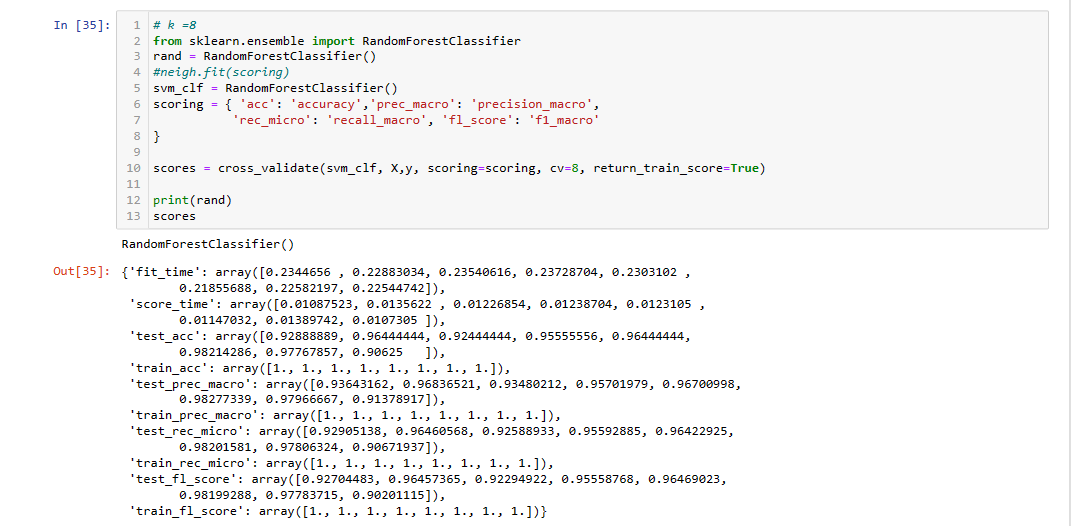
******

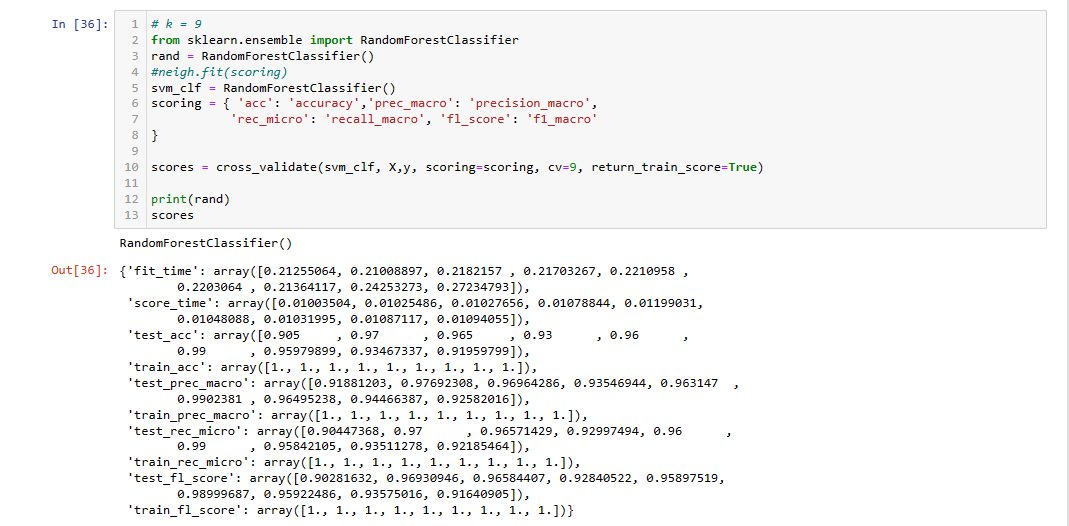
******

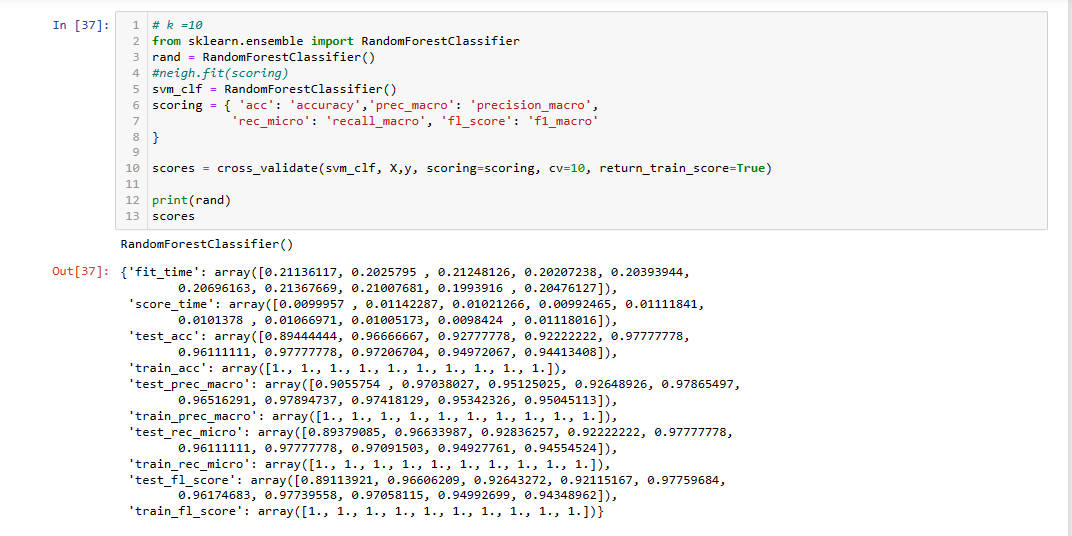
******

******

******

******

******

******

1. Từ kết quả ở câu trên lập bảng tổng hợp kết quả so sánh kết quả huấn luyện trên tập huấn luyện và kiểm tra từ câu trên (với các độ đo accuracy, recall, f1-score, và thời gian thực thi), nhận xét kết quả thay đổi thế nào với số k, và thử giải thích kết quả.

***Bài làm***

***SVM:***

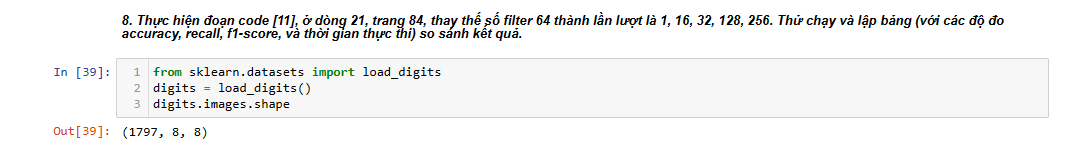
* Kết quả huấn luyện đạt khá cao theo từng k, đối với từng k chỉ có chêch lệch không đáng kể. Thời gian thực thi nằm trong khoản chấp nhận, tùy thuộc vào cấu hình máy thực nghiệm mà có thời gian khác nhau.
* Đối với từng k thì accuracy có xu hướng giảm dần từ k=3 đến k =10, trong khi f1\_score (train) vẫn không đổi và ngược lại với tập test có xu hướng giảm dần

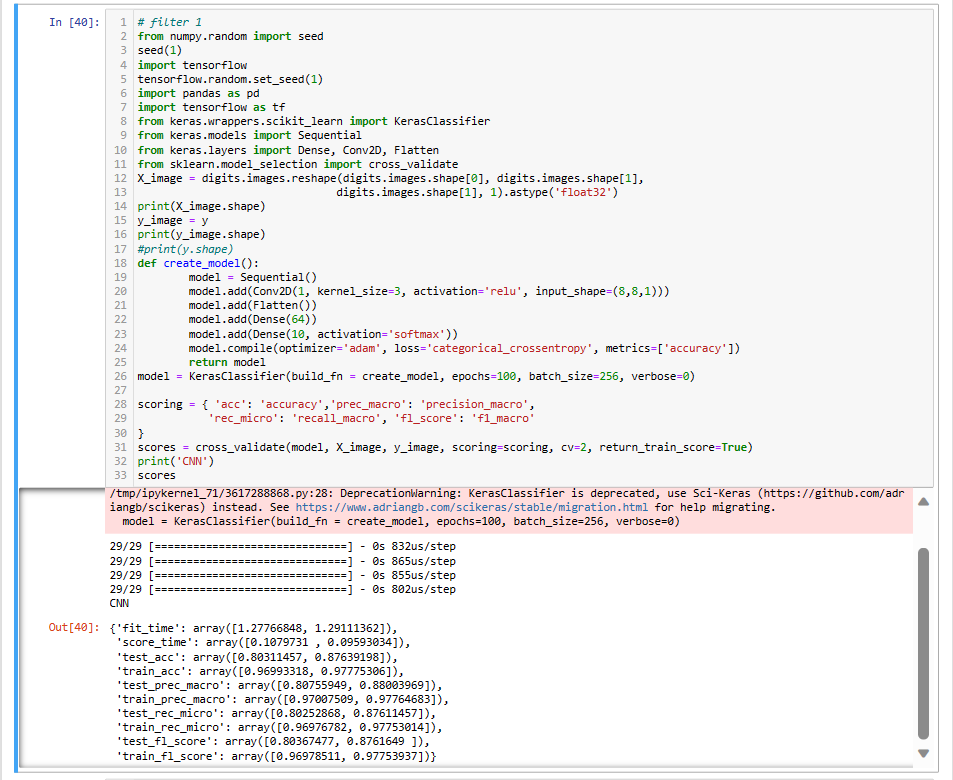
***Random Forest:***

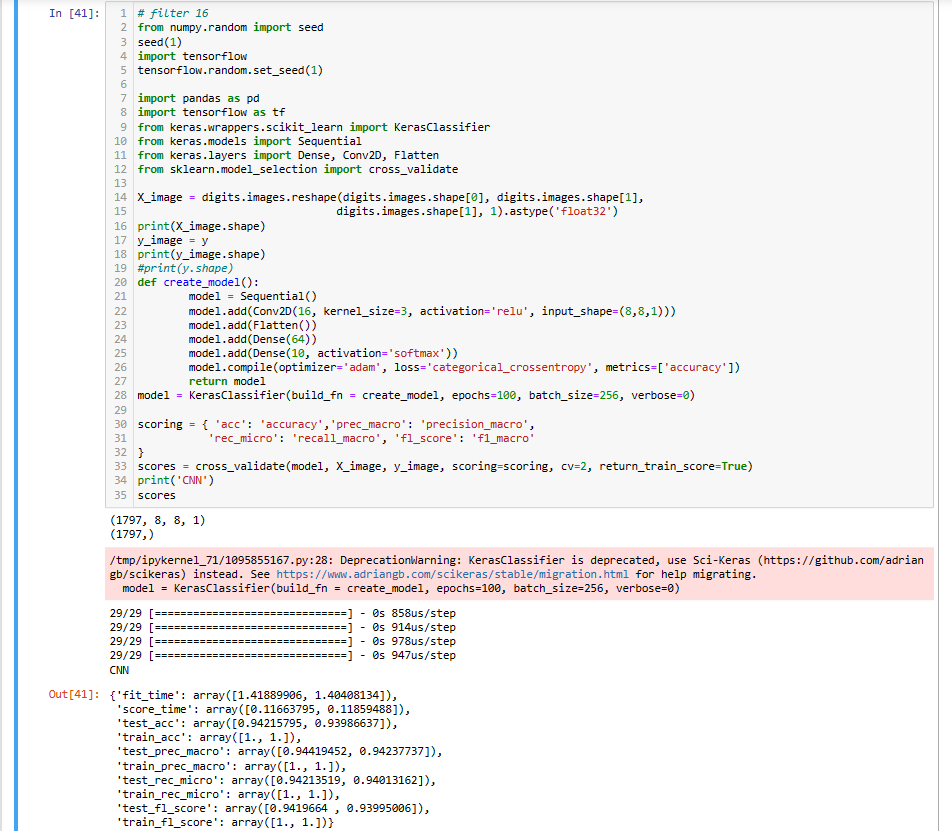
* Kết quả huấn luyện đạt khá cao theo từng k, đối với từng k chỉ có chêch lệch không đáng kể. Thời gian thực thi nằm trong khoản chấp nhận, tùy thuộc vào cấu hình máy thực nghiệm mà có thời gian khác nhau.
* Đối với từng k thì accuracy có xu hướng giảm dần từ k=3 đến k =10, trong khi f1\_score (train) vẫn không đổi ở mức 1 và ngược lại với tập test có xu hướng giảm không liên tục.

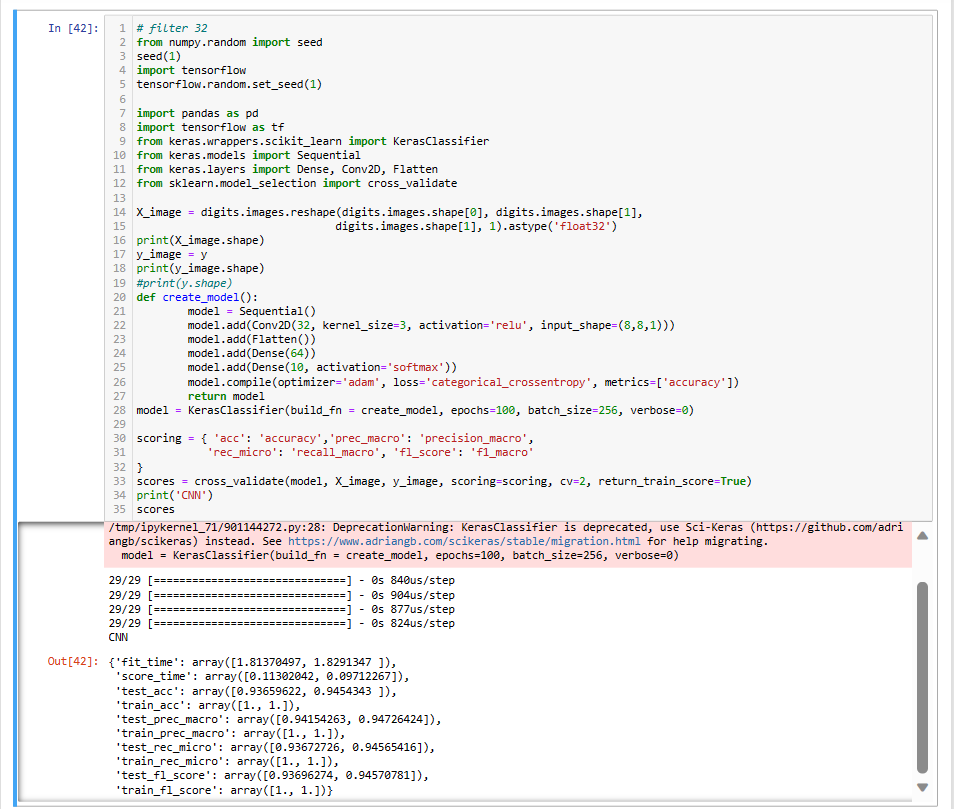
1. Thực hiện đoạn code [11], ở dòng 21, trang 84, thay thế số filter 64 thành lần lượt là 1, 16, 32, 128, 256. Thử chạy và lập bảng (với các độ đo accuracy, recall, f1-score, và thời gian thực thi) so sánh kết quả.

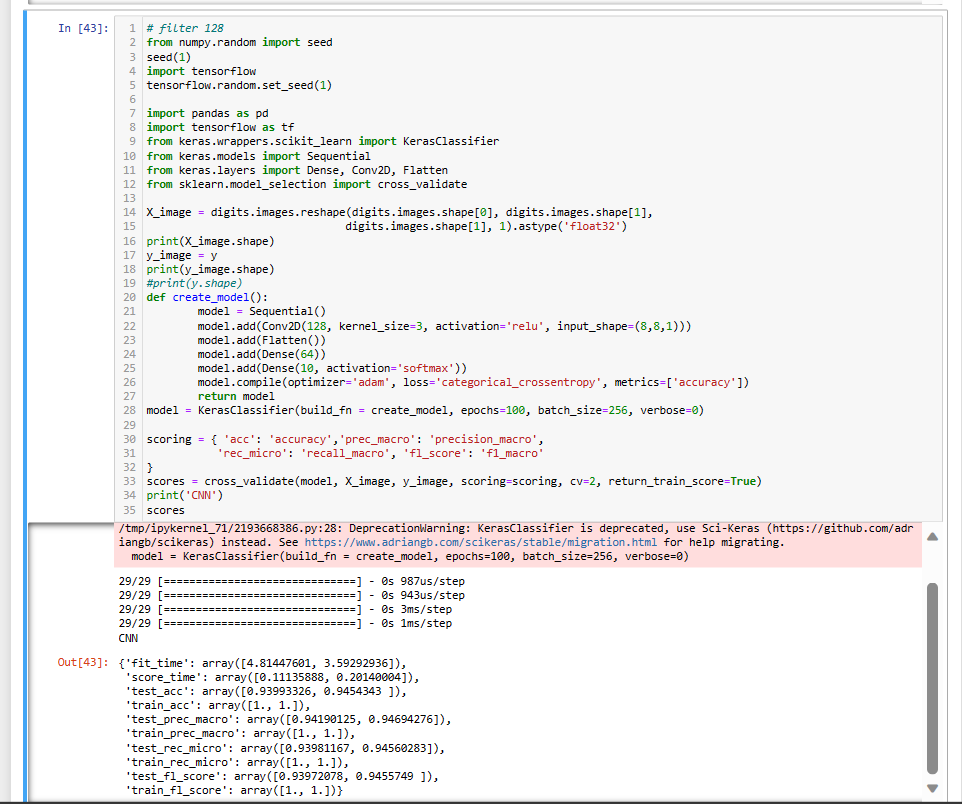
***Bài làm***

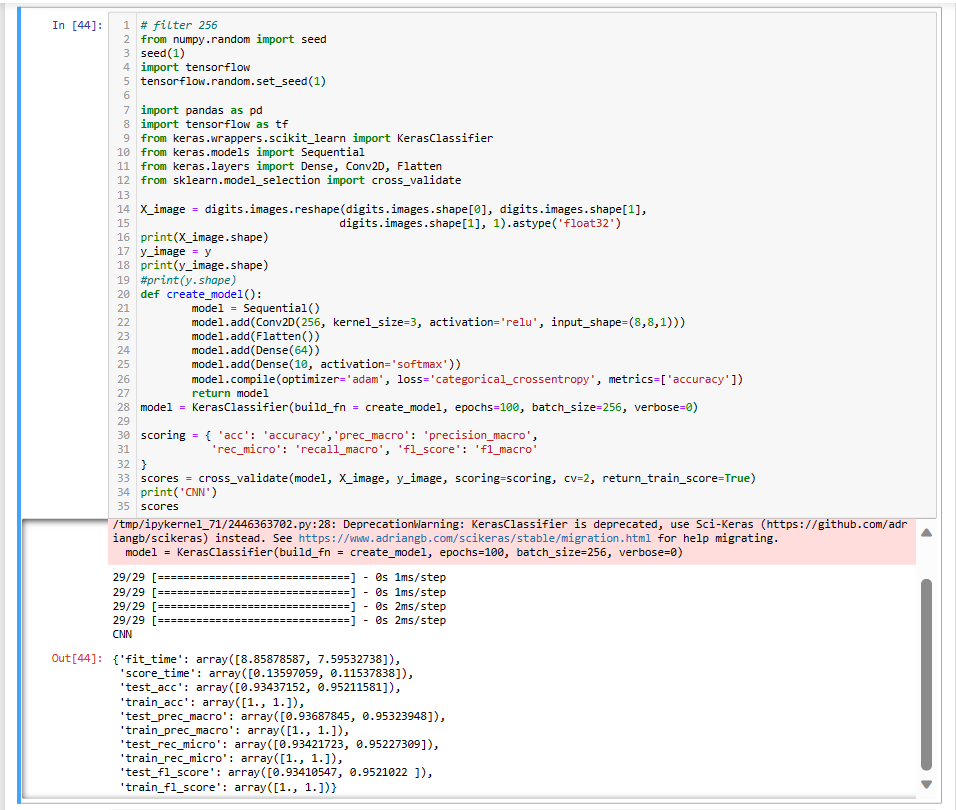
**

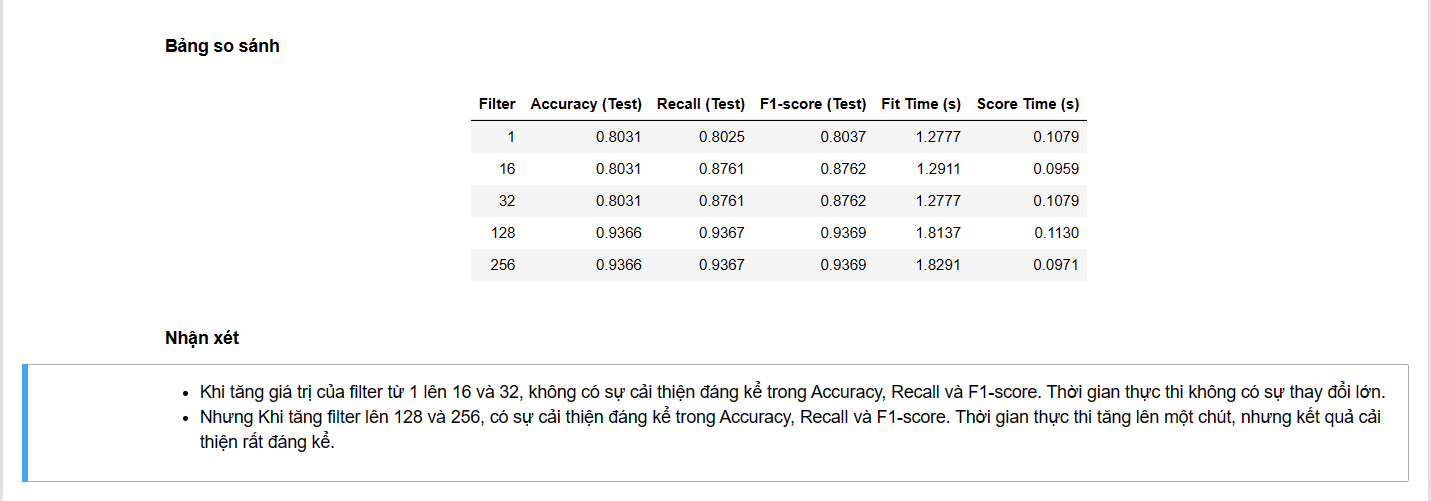
**

**

**

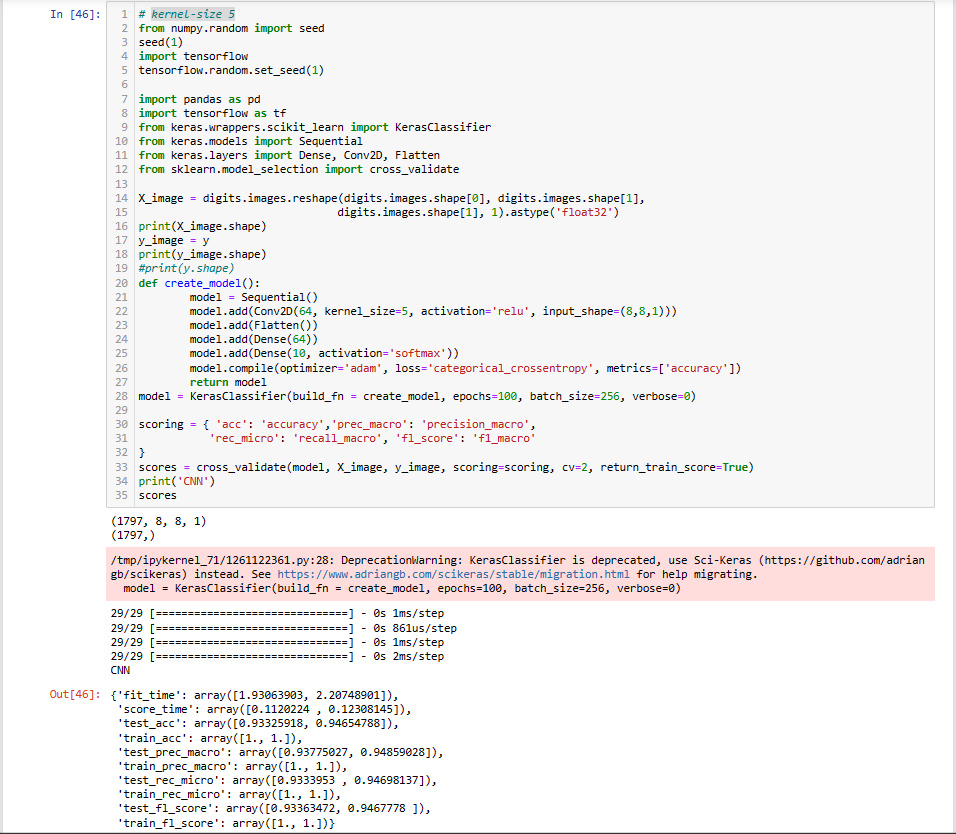
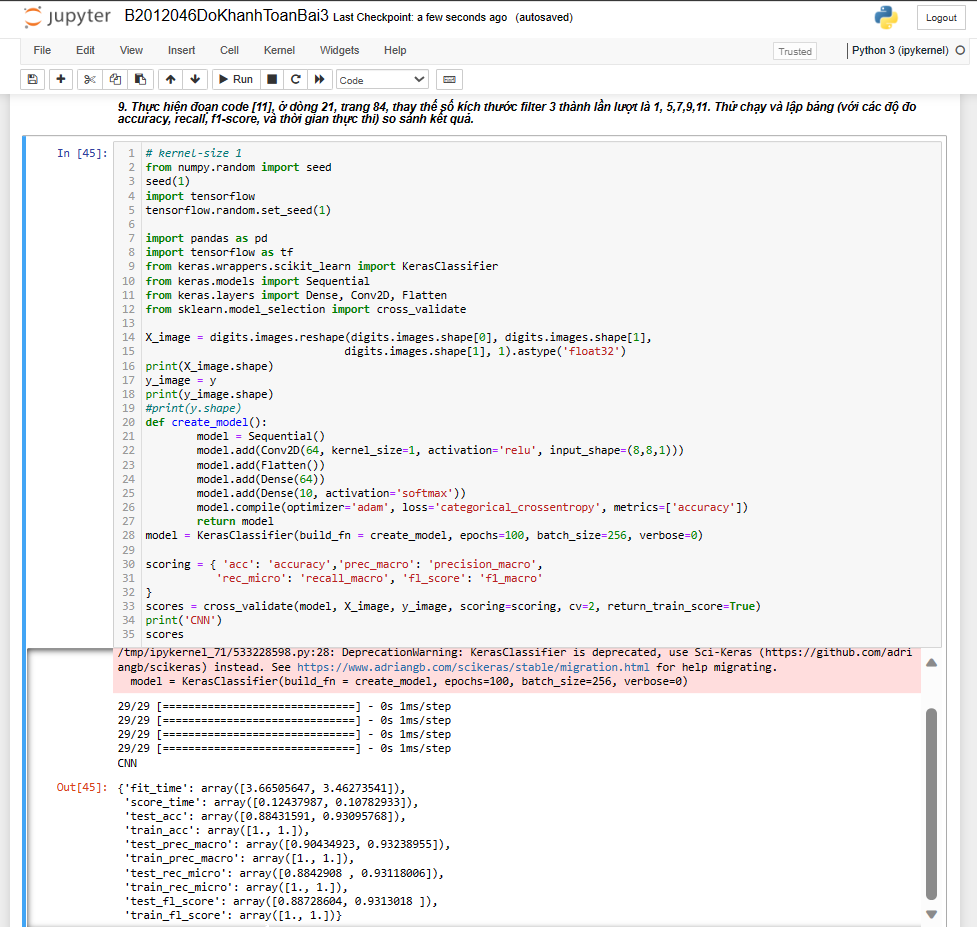
**

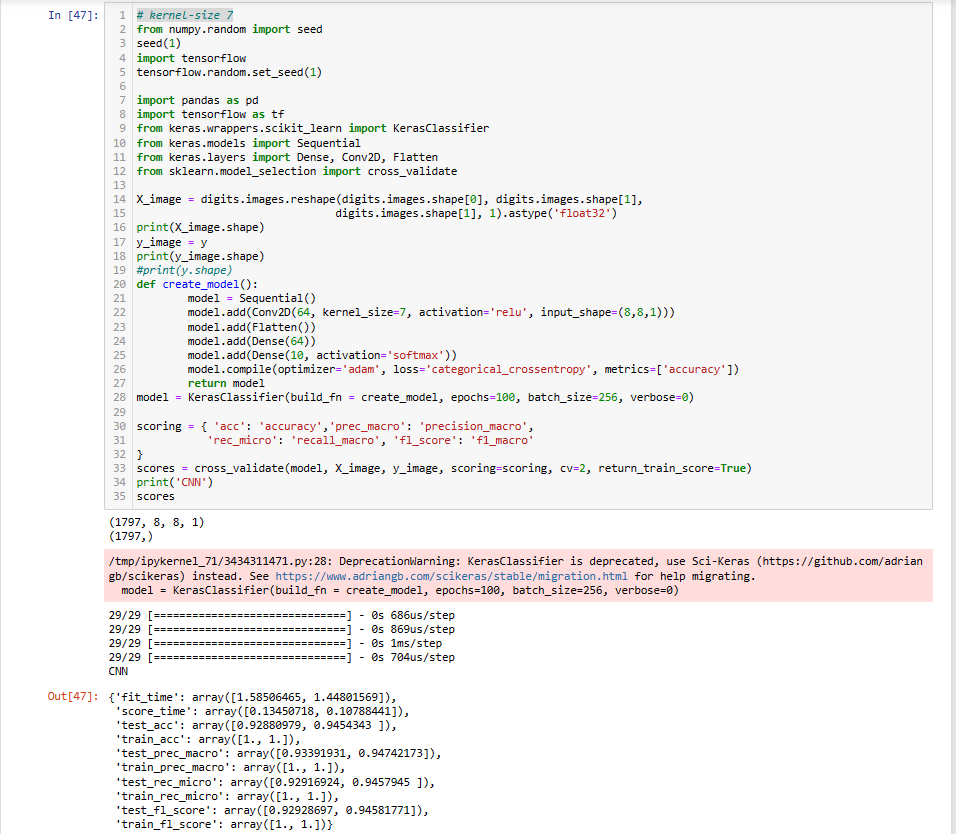
**

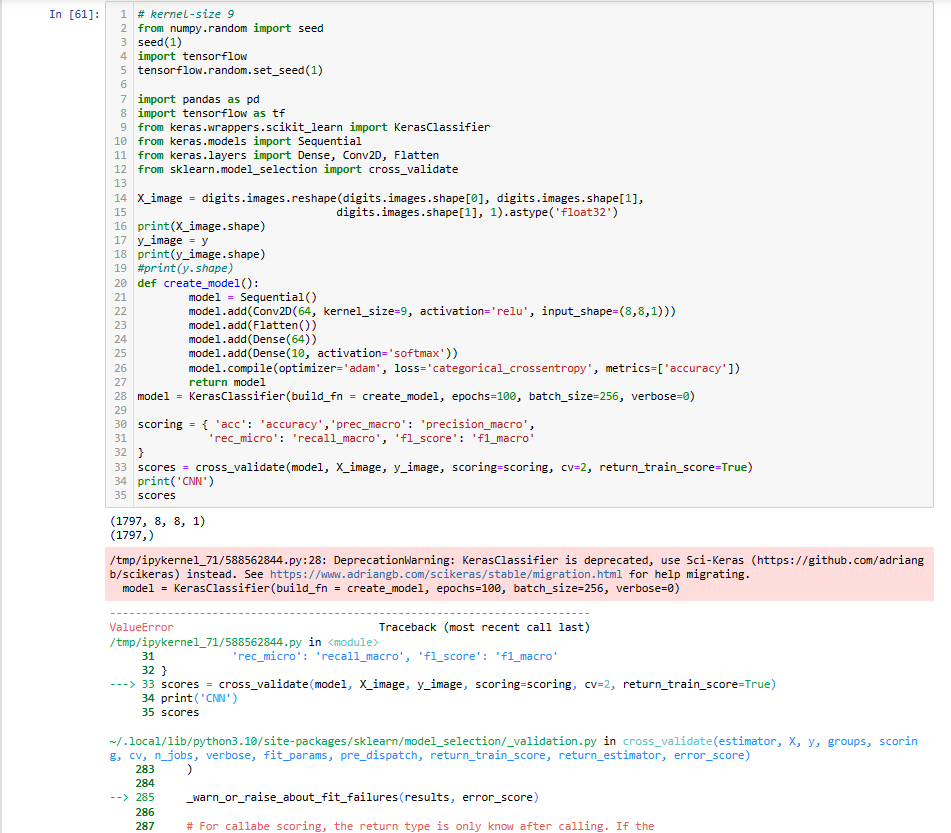
*SO SÁNH:  
  
*

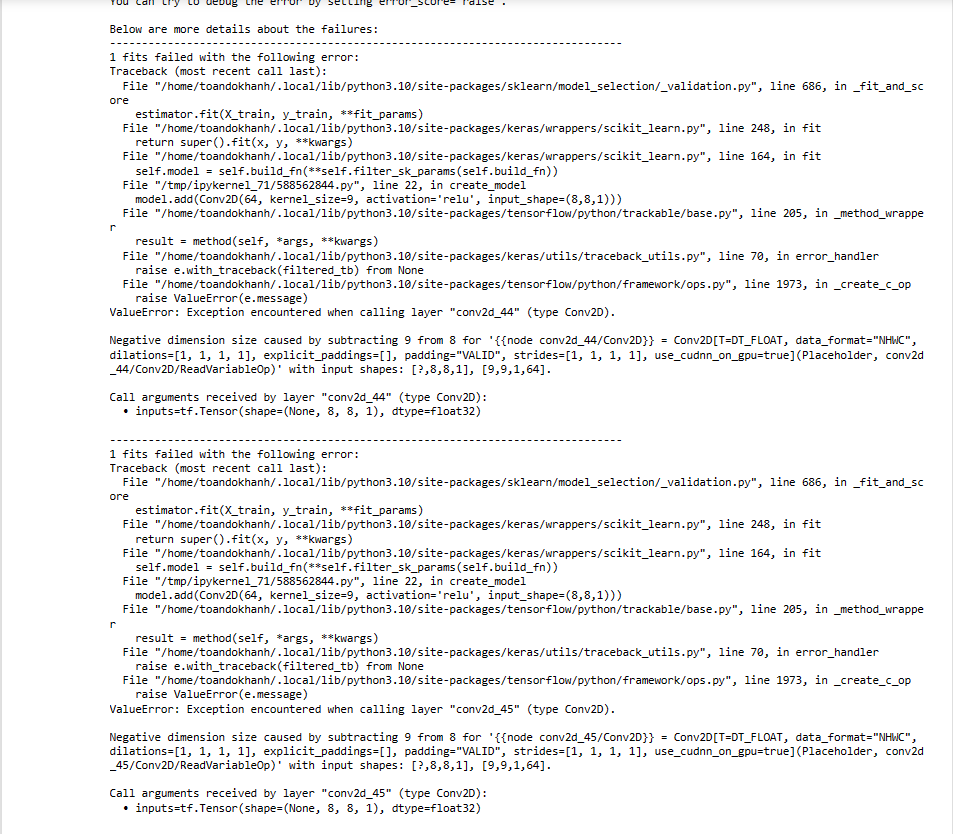
1. Thực hiện đoạn code [11], ở dòng 21, trang 84, thay thế số kích thước filter 3 thành lần lượt là 1, 5,7,9,11. Thử chạy và lập bảng (với các độ đo accuracy, recall, f1-score, và thời gian thực thi) so sánh kết quả.

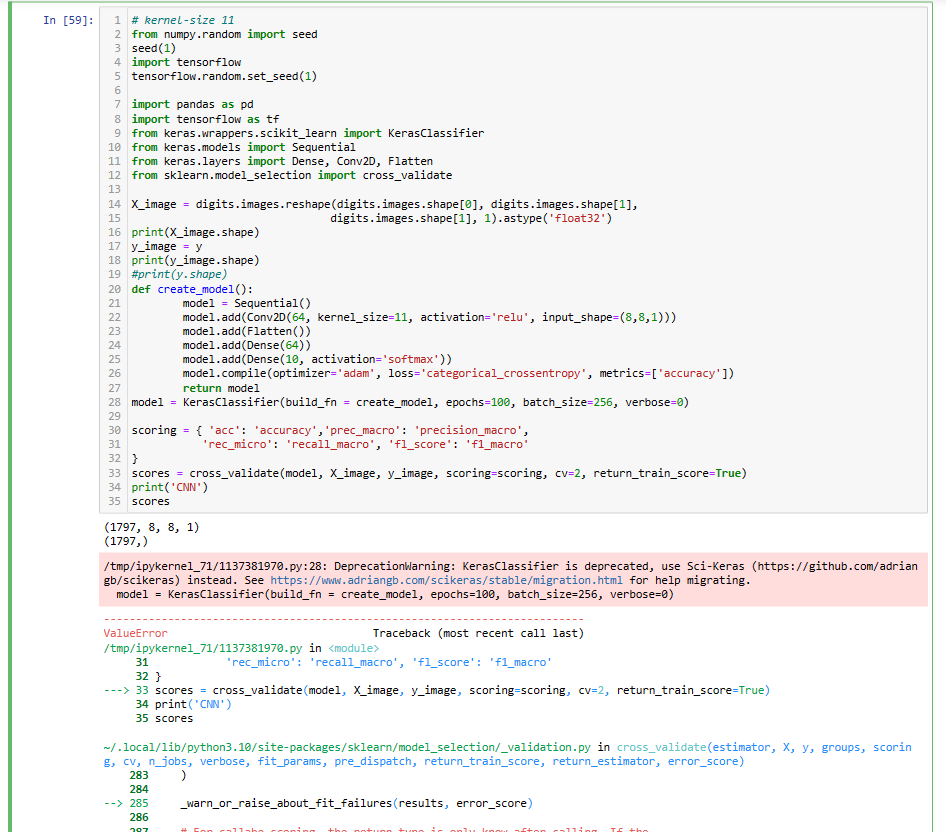
***Bài làm***

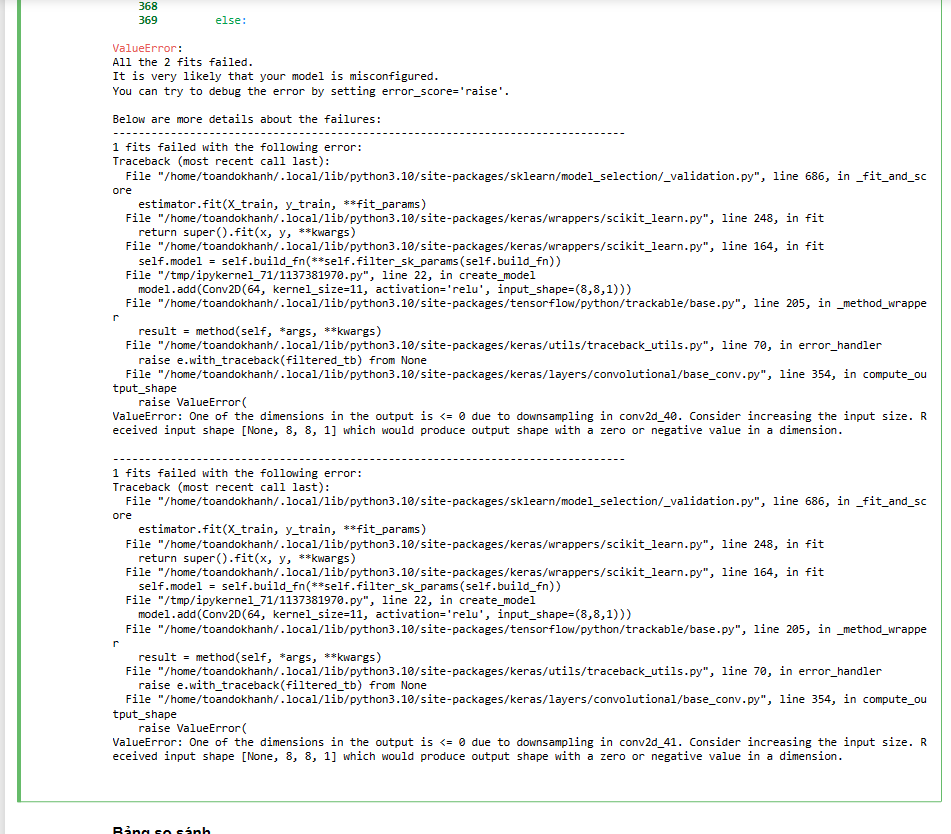
******

******

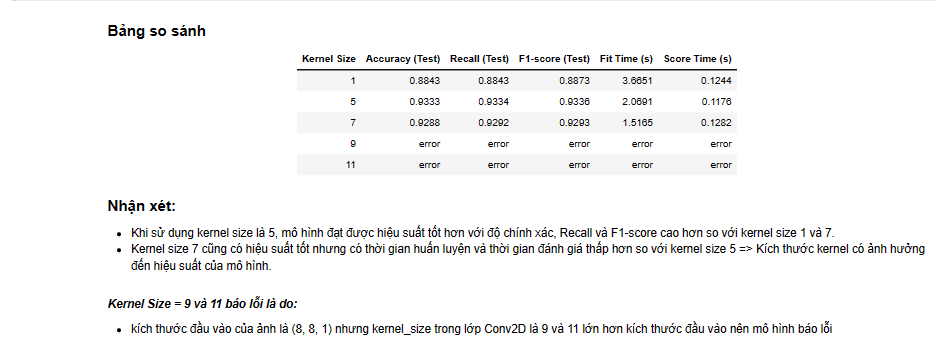
******

******

******

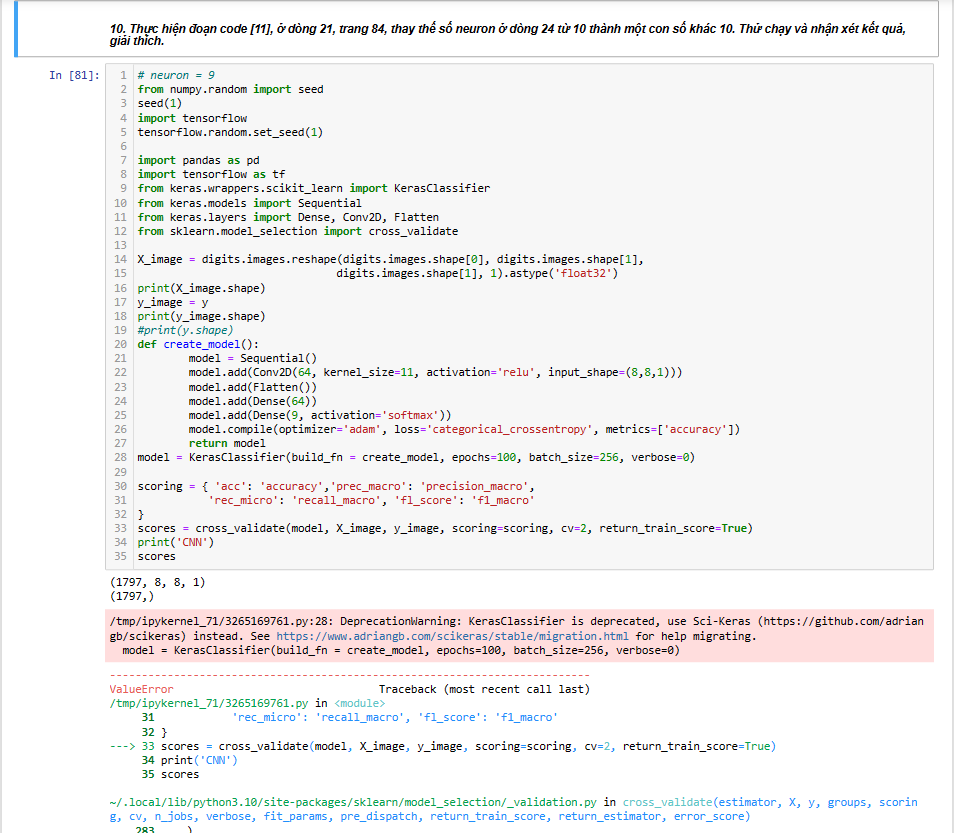
******

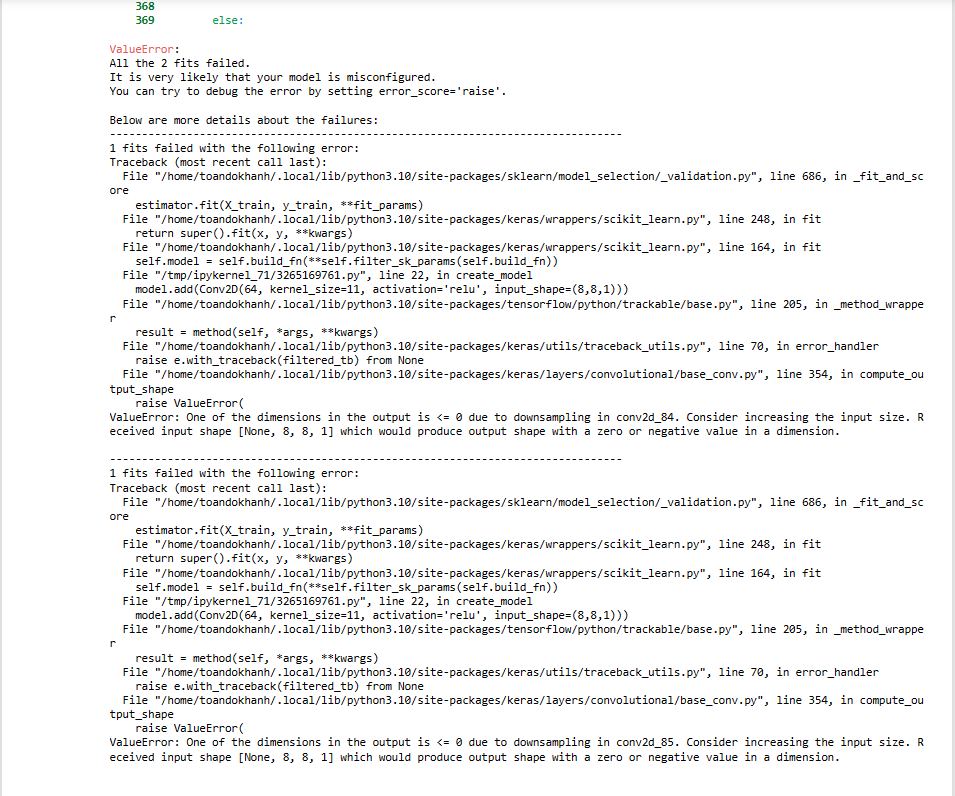
***SO SÁNH***

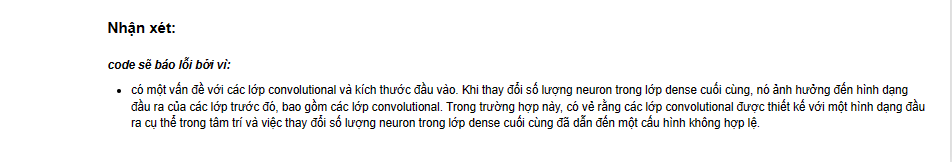
******

1. Thực hiện đoạn code [11], ở dòng 21, trang 84, thay thế số neuron ở dòng 24 từ 10 thành một con số khác 10. Thử chạy và nhận xét kết quả, giải thích.

***Bài làm***

******

******

******

******

**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)