

# Report Guides for Python for Data Analysis

Quách Đình Hoàng

2022/04/05

**Chú ý:** Việc tuân theo hướng dẫn này không nhất thiết giúp các nhóm được đánh giá cao (vì nó còn phụ thuộc vào chất lượng công việc được thực hiện) nhưng nó sẽ giúp các nhóm tổ chức báo cáo và hoàn thành project.

## 1. Tóm tắt (abstract) - 1 đoạn, khoảng 200-300 từ

Phần này là một đoạn mô tả động lực (motivation) của các câu hỏi nghiên cứu được đề cập trong project, các phương pháp nhóm sử dụng, và kết quả.

## 2. Giới thiệu (introduction) - Khoảng 1/2 đến 1 trang

Phần này giải thích các câu hỏi nghiên cứu và tại sao nó quan trọng. Nhóm nên cung cấp một số thông tin về bối cảnh liên quan đến các câu hỏi nghiên cứu. Những gì là biến giải thích (input) và biến phản hồi (output) của bài toán cần được làm rõ. Chẳng hạn: “Input bài toán là tập các {hồ sơ cho vay, lịch sử giá của một cổ phiếu, ...}. Chúng tôi sử dụng các thuật toán {linear regression, logistic regression, ...} để dự đoán {khả năng hồ sơ được duyệt cho vay, giá cổ phiếu sau một tháng, ...}”. Mỗi nhóm có thể có input/output rất khác nhau. Việc xác định rõ input/output giúp người đọc dễ theo dõi phần còn lại của báo cáo.

## 3. Dữ liệu (data) - khoảng 1/2 đến 1 trang

Phần này mô tả tập dữ liệu nhóm sử dụng (nó đến từ đâu, được thu thập như thế nào, các đối tượng là gì, các đặc trưng/thuộc tính là gì, ...). Phần này không nên quá 1 trang. Tập dữ liệu có bao nhiêu instance/example? In ra vài instance trong tập dữ liệu nếu cần thiết. Nhóm có thực hiện tiền xử lý không? Nếu có thì như thế nào? (ví dụ: cách nhóm xử lý missing values, chuẩn hóa các thuộc tính, ...).

## 4. Trực quan hóa dữ liệu (data visualization) - khoảng 1-2 trang

Ở phần này, nhóm cần thực hiện phân tích dữ liệu sơ bộ, bao gồm một số thống kê tóm tắt và biểu đồ trực quan hóa, cùng với một số giải thích về cách chúng giúp nhóm hiểu thêm về tập dữ liệu. Các phân tích này nên hướng đến các câu hỏi nghiên cứu (bài toán) mà nhóm đặt ra trong đề xuất.

## 5. Mô hình hóa dữ liệu (data modeling) - khoảng 1-2 trang

Ở phần này, nhóm cần mô tả vắn tắt các phương pháp mà nhóm sử dụng trong việc trả lời các câu hỏi đặt ra. Nhóm cần mô tả các thủ tục thống kê (nếu có) và tóm tắt ngắn gọn ý tưởng của các thuật toán học máy mà nhóm sử dụng. Nhóm nên sử dụng các công thức hoặc hình ảnh để hỗ trợ phần giải thích ý tưởng của các thuật toán. Ví dụ: nếu nhóm sử dụng thuật toán logistic regression thì nhóm nên mô tả định nghĩa của hàm logistic/softmax. Nhóm cần thể hiện là mình hiểu cách thuật toán hoạt động. Mặc dù GV có thể biết/hiểu những thuật toán nhóm sử dụng, những người đọc khác có thể không. Do đó, với mỗi thuật toán, nhóm cần dành một đoạn văn mô tả cách nó hoạt động. Điều này đặc biệt cần thiết khi nhóm dùng các thuật toán chưa được trình bày trong môn học (chẳng hạn: decision tree, SVM, neural network, ...), cho bài toán của mình.

## **6. Thực nghiệm, kết quả, và thảo luận (experiments, results, and discussions) - khoảng 1-3 trang**

Nhóm nên mô tả chi tiết về các câu hỏi nghiên cứu và các phương pháp, thuật toán nhóm sử dụng để trả lời chúng. Để trình bày kết quả, nhóm nên dùng kết hợp các bảng và biểu đồ. Nếu bài toán là phân loại thì nhóm cần mô tả confusion matrix và các độ đo accuracy, precision, recall, và F-measure. Nếu bài toán là hồi quy, nhóm cần mô tả RMSE, R-square. Nhóm cũng cần trình bày các biểu đồ trực quan về kết quả. Nhóm cần đưa ra các nhận xét, bình luận về các số liệu, bảng biểu, biểu đồ. Các bảng, biểu đồ của nhóm phải được đánh số, đặt tiêu đề, thêm các chú giải (legends, axis labels), và có kích thước, phông chữ dễ đọc khi in.

## **7. Kết luận (conclusions) - khoảng 1-2 đoạn**

Phần này tóm tắt lại những kết quả chính của nhóm. Phát hiện nào thú vị và gây ngạc nhiên nhất? Nếu nhóm có nhiều thời gian hơn, nhóm sẽ làm thêm cái gì?

## **8. Phụ lục (appendices) - tối đa 2 trang**

Phần này là tùy chọn. Nó bao gồm các mô tả chi tiết hơn về các thuật toán hoặc các vấn đề kỹ thuật mà nhóm không có đủ không gian để trình bày ở các mục trên.

## **9. Đóng góp (contributions) - không có giới hạn**

Phần này mô tả những gì mà mỗi thành viên trong nhóm đóng góp vào project này.

## **10. Tham khảo (references) - không có giới hạn**

Phần này mô tả các tài liệu nhóm đã tham khảo để hoàn thành báo cáo này. Các tài liệu này có thể là (1) các bài báo, sách, blog mà nhóm đã tham khảo, (2) code, library nhóm đã sử dụng (ví dụ: pandas, matplotlib, scikit-learn, ...). Tài liệu tham khảo cần được viết theo đúng một trong các định dạng MLA, APA, hoặc IEEE. Nếu nhóm không biết các định dạng này thì mỗi tài liệu tham khảo phải bao gồm (theo thứ tự): các tác giả (cách nhau bằng dấu phẩy), tựa đề bài báo, tên conference/journal, nhà xuất bản, năm.