

# Project Guides for Data Mining

Quách Đình Hoàng

2021/11/01

## Thông tin chung về project môn học

- Project môn học là một case study về việc áp dụng data mining dùng R trên tập dữ liệu do mỗi nhóm lựa chọn. Tập dữ liệu có thể đã tồn tại hoặc mỗi nhóm có thể tự thu thập dữ liệu bằng kỹ thuật web scraping. Mỗi nhóm có thể chọn dữ liệu dựa trên mối quan tâm của nhóm. Mục tiêu của project này là để các nhóm thể hiện sự thành thạo trong các kỹ thuật đã được đề cập trong môn học và áp dụng chúng vào tập dữ liệu mới theo cách có ý nghĩa.
- Mục tiêu của project là để GV biết rằng nhóm thành thạo trong việc áp dụng data mining với R, và các thư viện liên quan để giải quyết một bài toán cụ thể trên một tập dữ liệu cụ thể. Nhóm cần thể hiện sự thành thạo trong việc thao tác trên dữ liệu, trực quan hóa dữ liệu, phân tích dữ liệu để đưa ra dự đoán, diễn giải, và trình bày kết quả.
- Mỗi nhóm nên tạo một số loại biểu đồ trực quan thú vị về tập dữ liệu đã lựa chọn dùng R. Không có giới hạn về công cụ hoặc thư viện nào mỗi nhóm có thể sử dụng, nhưng các nhóm ít nhất phải sử dụng các thư viện `dplyr`, `ggplot2`. Các nhóm không cần phải trực quan hóa tất cả dữ liệu cùng một lúc. Một biểu đồ chất lượng cao sẽ nhận được đánh giá cao hơn nhiều so với nhiều biểu đồ chất lượng kém. Chất lượng bài thuyết trình cũng quan trọng. Bài trình bày súc tích và rõ ràng sẽ được đánh giá cao. Tất cả các phân tích phải được thực hiện bằng Rmarkdown với cú pháp markdown (.Rmd).
- **Teamwork:** Các thành viên phải hoàn thành project với tư cách là một nhóm. Tất cả các thành viên trong nhóm phải có trách nhiệm đóng góp như nhau (cả về chất lượng và số lượng) vào việc hoàn thành project. Các thành viên trong nhóm nên dành thời gian để làm việc cùng nhau. Bất kỳ ai bị đánh giá là không có đóng góp đủ cho sản phẩm cuối cùng sẽ bị phạt điểm. Mặc dù các thành viên trong nhóm có thể có nền tảng và phụ trách các công việc khác nhau, nhưng trách nhiệm của mọi thành viên trong nhóm là phải hiểu tất cả những gì được thực hiện trong project và lý giải được cách tiếp cận trong project là hợp lý.

## Dữ liệu (Data)

- Để đạt được kết quả tốt nhất với project môn học, điều quan trọng là các nhóm phải chọn một tập dữ liệu mà nhóm có thể quản lý được. Tập dữ liệu của mỗi nhóm phải dễ dàng truy cập, có ít nhất 1000 quan sát, tối thiểu 8 biến, và nên ở dạng csv (có thể có ngoại lệ nhưng nhóm phải trao đổi với GV trước). Các biến trong tập dữ liệu nên bao gồm các biến phân loại, các biến số rời rạc và các biến số liên tục.
- Dưới đây là danh sách các nguồn dữ liệu mà các nhóm có thể xem xét để lựa chọn. Các nhóm không bị giới hạn ở những nguồn dữ liệu này. Nhưng đây là các nguồn hữu ích để các nhóm có thể tìm thấy tập dữ liệu phù hợp cho vấn đề mà nhóm quan tâm:
  - Kaggle datasets: <https://www.kaggle.com/datasets>
  - Awesome public datasets: <https://github.com/awesomedata/awesome-public-datasets>
  - UC Irvine machine learning datasets: <https://archive.ics.uci.edu/ml/index.php>

## Những thứ cần nộp

- **Proposal:** hạn chót: **23h59, 17/04/2022**
- **Milestone:** hạn chót: **23h59, 15/05/2022**
- **Presentation:** hạn chót: **23h59, 29/05/2022**
- **Report:** hạn chót: **23h59, 05/06/2022**

## Đề xuất (Proposal)

Đề xuất cần bao gồm các phần sau:

### Phần 1 - Giới thiệu

- Phần này nên giới thiệu câu hỏi nghiên cứu chung và dữ liệu của nhóm (nó đến từ đâu, thu thập như thế nào, các trường hợp xảy ra là gì, các biến số là gì, v.v.). Phần này không quá 1 trang (không kể số liệu).

### Phần 2 - Dữ liệu

- Đặt dữ liệu của nhóm vào thư mục **data**, cho biết số chiều (số dòng, số cột) và giải thích các biến trong file **README.md** trong thư mục đó. Thực hiện phần EDA để in ra vài dòng đầu và cuối trong tập dữ liệu của nhóm.

### Phần 3 - Kế hoạch phân tích dữ liệu

- Các biến kết quả (phản hồi, **Y**) và dự đoán (giải thích, **X**) mà nhóm sẽ sử dụng để trả lời câu hỏi của mình.
- 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. (Nhóm có thể cập nhật những thông tin khi hoàn thành project.)
- (Các) phương pháp mà nhóm tin rằng sẽ hữu ích trong việc trả lời (các) câu hỏi đặt ra. (Nhóm có thể cập nhật những thông tin này khi hoàn thành project.)
- Kế hoạch thực hiện và phân công việc.

## Bài thuyết trình (Presentation)

- Mỗi nhóm có **tối đa 10 phút** để trình bày và mỗi thành viên trong nhóm nên nói điều gì đó quan trọng. Các nhóm có thể trình bày trực tiếp trong buổi báo cáo của mình hoặc ghi âm trước và gửi video để phát trong buổi báo cáo. Ngay cả khi đã ghi video trước, mỗi nhóm vẫn phải có mặt ở buổi báo cáo để xử lý các vấn đề phát sinh và trả lời các câu hỏi.
- Mỗi nhóm cần chuẩn bị một bản trình chiếu (slides) bằng Rmarkdown hoặc Powerpoint. Không có giới hạn về số lượng slide trình bày mỗi nhóm có thể sử dụng, chỉ có giới hạn thời gian (tối đa 10 phút). Mỗi thành viên trong nhóm nên có cơ hội phát biểu trong buổi thuyết trình. Bản trình bày của nhóm không nên chỉ là tất cả những gì nhóm đã thử (chúng tôi đã làm điều này, sau đó chúng tôi đã làm điều này, v.v.), thay vào đó nó phải truyền tải những lựa chọn mà nhóm đã thực hiện, tại sao và những gì nhóm thấy thú vị.
- Lịch trình thuyết trình: Các bài thuyết trình sẽ diễn ra trong buổi báo cáo cuối cùng. Mỗi nhóm có thể chọn trình bày trực tiếp bài trình bày của mình hoặc ghi âm trước. Trong buổi báo cáo, mỗi nhóm sẽ xem các bài thuyết trình từ các nhóm khác và cung cấp phản hồi dưới hình thức đặt câu hỏi và đánh giá ngang hàng. Thứ tự trình bày sẽ được GV lựa chọn ngẫu nhiên.

## Báo cáo (Report)

- Cùng với các slide trình bày, mỗi nhóm cần cung cấp một báo cáo về project của nhóm. Báo cáo này cũng được viết dùng Rmarkdown. Báo cáo có định dạng là `.html/.pdf` được xuất ra từ `.Rmd`.
- Báo cáo này là bản tóm tắt về project của nhóm. Nó cung cấp thông tin về tập dữ liệu nhóm đang sử dụng, bài toán nhóm muốn giải quyết, các phương pháp nhóm đã thử nghiệm cùng kết quả và nhận xét.

## Tổ chức project

Project của nhóm cần được tổ chức theo cấu trúc sau:

- **Thư mục data:** gồm tập dữ liệu (dataset) ở dạng `.csv` được sử dụng cùng với file `README.md` để mô tả về tập dữ liệu.
- **Thư mục image (nếu có):** gồm các file ảnh được sử dụng.
- **Đề xuất:** gồm `Proposal.Rmd` và `Proposal.html`.
- **Bài thuyết trình:** gồm `Presentation.Rmd` và `Presentation.html`.
- **Báo cáo:** gồm `Report.Rmd` và `Report.html/Report.pdf`.

**Ghi chú:** Cách trình bày (Proposal, Presentation, Report) cũng được tính điểm. Các nhóm nên trình bày chúng sao cho rõ ràng. Code cũng nên được định dạng hợp lý.

## Cách tính điểm

- **Proposal:** 10%
- **Milestone:** 10%
- **Presentation:** 20%
- **Report:** 40%
- **Peer assessment:** 10%
- **Q&A:** 10%

## Các tiêu chí chấm điểm

- **Nội dung:** Mức độ liên quan của dữ liệu với bài toán đặt ra, độ khó của bài toán, và chất lượng của dự án.
- **Tính đúng đắn:** Các phương pháp có được thực hiện và giải thích một cách chính xác không?
- **Viết và trình bày:** Chất lượng của bài trình bày, của báo cáo, và giải thích.
- **Sự sáng tạo và tư duy phản biện:** Project có được suy nghĩ cẩn thận không? Những hạn chế có được xem xét cẩn thận? Thời gian và công sức đã dành cho việc lập kế hoạch và thực hiện dự án?

## Đánh giá giữa các thành viên trong nhóm (peer assessment)

- Mỗi thành viên sẽ được yêu cầu cung cấp một đánh giá cá nhân trong đó mỗi bạn đánh giá sự đóng góp và tinh thần đồng đội của từng thành viên trong nhóm trên thang điểm 10. Mỗi thành viên cần báo cáo những công việc mà các thành viên khác đã thực hiện và tỷ lệ phần trăm đóng góp của mỗi thành viên trong nhóm. Đánh giá cá nhân là điều kiện tiên quyết để nhận được điểm cho project. Nếu bạn cho rằng một cá nhân nào trong nhóm đã làm ít hơn 20% công việc, hãy cung cấp một số lời giải thích. Nếu bất kỳ cá nhân nào nhận được điểm đánh giá trung bình cho thấy rằng họ đã làm ít hơn 15% công việc, người này sẽ nhận được một nửa số điểm của những người còn lại trong nhóm. Đánh

giá của bạn trong nhóm chiếm 10% điểm project. Nói chung, giáo viên sẽ tôn trọng ý kiến của các thành viên trong nhóm về mức độ đóng góp. Tuy nhiên, nếu có những điểm mà giáo viên thấy chưa hợp lý, đánh giá này có thể được xem xét và điều chỉnh bởi giáo viên.