


THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/k-WChA-UYhE>
- Link slides (dạng .pdf đặt trên Github):
<https://github.com/sinhlh14/CS2205.APR2023/blob/main/Sinh%20Le%CC%82%20Hoa%CC%80ng%20-%20xCS2205.DeCuong.FinalReport.Slide.pdf>
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none">• Họ và Tên: Lê Hoàng Sinh• MSSV: CH1901028 	<ul style="list-style-type: none">• Lớp: CS2205.APR2023• Tự đánh giá (điểm tổng kết môn): 8.0/10• Số buổi vắng: 0• Tỷ lệ câu hỏi QT cá nhân đã làm: 100%• Tỷ lệ câu hỏi QT đã làm của nhóm: 100%• Link Github: https://github.com/sinhlh14/CS2205.APR2023• Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none">○ Lên ý tưởng○ Làm báo cáo○ Làm poster○ Làm slide○ Làm video YouTube
---	---

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

LỰA CHỌN ĐẶC TRƯNG CHO MÔ HÌNH TUYẾN TÍNH SỬ DỤNG CRITERION RELAXATION

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

FEATURE SELECTION USING CRITERION RELAXATION

TÓM TẮT (*Tối đa 400 từ*)

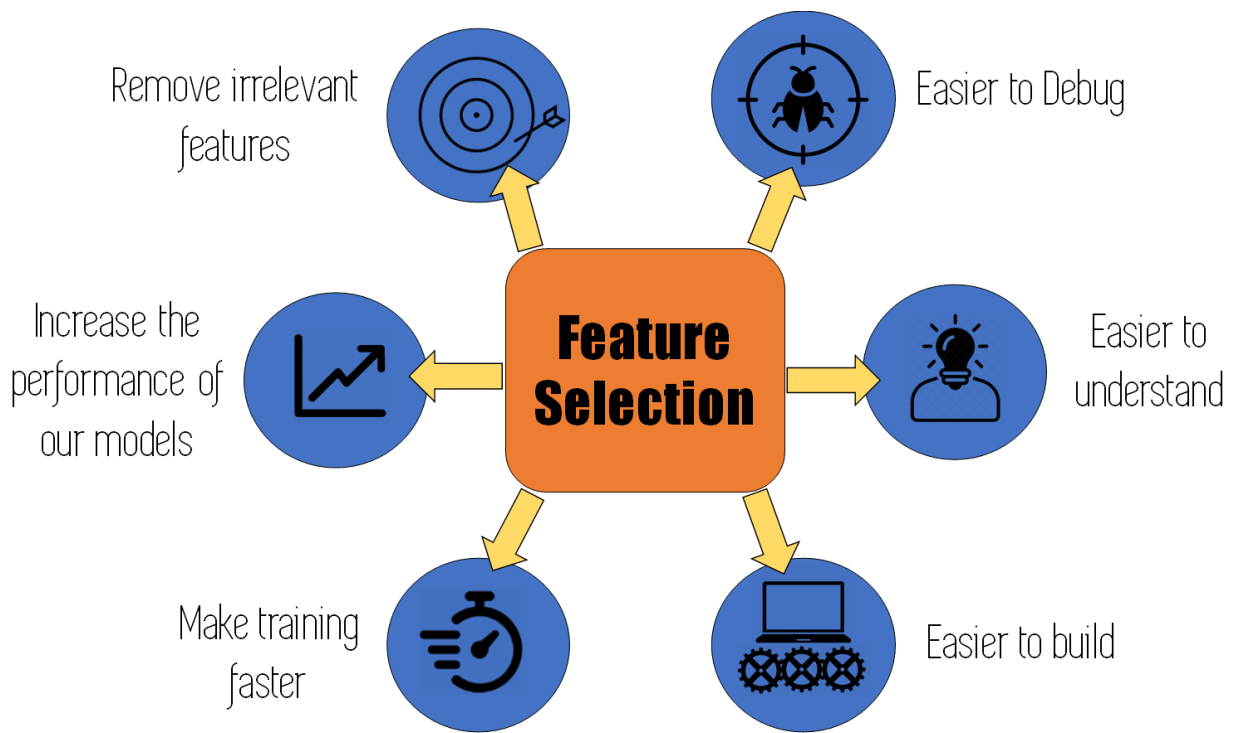
Trong thời đại dữ liệu lớn này, các kỹ thuật lựa chọn đặc trưng đã được chứng minh từ lâu rằng chúng có thể đơn giản hóa mô hình, làm cho mô hình dễ hiểu hơn và tăng tốc quá trình học, đã trở nên ngày càng quan trọng. Trong số nhiều phương pháp được phát triển, lựa chọn đặc trưng theo chiều thuận (forward selection), chiều ngược (backward selection) và theo bước (stepwise selection) vẫn được sử dụng rộng rãi nhờ tính đơn giản và hiệu quả của chúng. Tuy nhiên, những phương pháp này còn nhược điểm khi đối mặt với các tập dữ liệu lớn, điều này dẫn đến song song hóa trong lựa chọn đặc trưng. Tuy nhiên, trong bài báo này, chúng tôi phân tích vấn đề tiềm ẩn với lựa chọn đặc trưng song song, và từ đó, chỉ ra rằng cần cải thiện tốc độ của lựa chọn đặc trưng theo chiều thuận và theo bước là cần thiết. Điều này dẫn chúng tôi đề xuất *criterion relaxation*, với ý tưởng thay thế một tiêu chí cho lựa chọn đặc trưng bằng một tiêu chí có liên quan mật thiết và đơn giản hơn trong giai đoạn đầu của việc lựa chọn đặc trưng, trước khi quay trở lại tiêu chí bình thường. Chúng tôi cũng tiến hành các thí nghiệm để chứng minh rằng các bước lựa chọn đặc trưng theo chiều ngược hiếm khi được tiến hành trong lựa chọn đặc trưng theo bước. Chúng tôi minh họa rằng khi số lượng đặc trưng vượt quá đáng kể số lượng mẫu, việc lựa chọn đặc trưng được ưu tiên hơn việc giảm chiều sử dụng PCA, vì PCA có thể gặp khó khăn trong việc ước lượng kém của covariance matrix.

GIỚI THIỆU (Tối đa 1 trang A4)

Lựa chọn đặc trưng là quá trình chọn ra một tập hợp con của các đặc trưng từ tập hợp ban đầu của các đặc trưng có sẵn trong dữ liệu. Mục tiêu của việc này là tìm ra các đặc trưng quan trọng nhất, có ảnh hưởng nhất đến mục tiêu hoặc có khả năng giải thích tốt nhất cho mô hình học máy. Việc lựa chọn đặc trưng giúp cải thiện hiệu suất và hiểu biết về mô hình, giảm thiểu sự quá khớp (overfitting) và tăng tốc quá trình học máy. Có nhiều phương pháp lựa chọn đặc trưng khác nhau, bao gồm tiêu chí thống kê, thuật toán học máy, kỹ thuật đặc trưng nhúng và kết hợp các phương pháp khác nhau để tạo ra tập hợp cuối cùng của các đặc trưng quan trọng.

Trong số các phương pháp trên, việc sử dụng các phương pháp học máy như forward selection, backward selection và stepwise selection vẫn được sử dụng rộng rãi cho các bài toán hồi quy tuyến tính. Việc lựa chọn đặc trưng theo các phương pháp này thì dễ hiểu và dễ thực hiện, chúng có thể được áp dụng một cách trực quan và có thể triển khai dễ dàng trong các ngôn ngữ lập trình thống kê phổ biến như R hay Python. Giúp giảm số lượng biến độc lập trong mô hình. Khi có một tập hợp lớn các biến, việc loại bỏ các biến không quan trọng hoặc trùng lặp giúp cải thiện khả năng diễn giải, giảm độ phức tạp của mô hình và giảm nguy cơ overfitting. Forward selection và stepwise selection vẫn giữ được hiệu suất của mô hình, đặc trưng sẽ được thêm vào hoặc loại bỏ một cách tuần tự giúp tiết kiệm thời gian tính toán so với việc kiểm tra tất cả các tập hợp con có thể của các biến độc lập. Ngoài ra, phương pháp này còn giúp tối ưu hóa tiêu chí đánh giá. Cả forward selection và stepwise selection có thể sử dụng các tiêu chí đánh giá như sai số dự đoán, độ chính xác, AIC, BIC, hay R-squared để tối ưu hóa mô hình, giúp tìm ra tập hợp biến độc lập tốt nhất dựa trên các tiêu chí này.

Bên cạnh những ưu điểm, vẫn tồn tại có những nhược điểm, đặc biệt với dữ liệu lớn, việc thử nghiệm tất cả các tổ hợp biến có thể trong forward selection và stepwise selection trở nên rất tốn thời gian tính toán. Để khắc phục điều này chúng tôi đề xuất criterion relaxation, với ý tưởng thay thế một tiêu chí lựa chọn đặc trưng bằng một tiêu chí khác có liên quan mật thiết và đơn giản hơn trong giai đoạn đầu của việc lựa chọn đặc trưng.



Đầu vào sẽ là tập dữ liệu gồm nhiều đặc trưng.

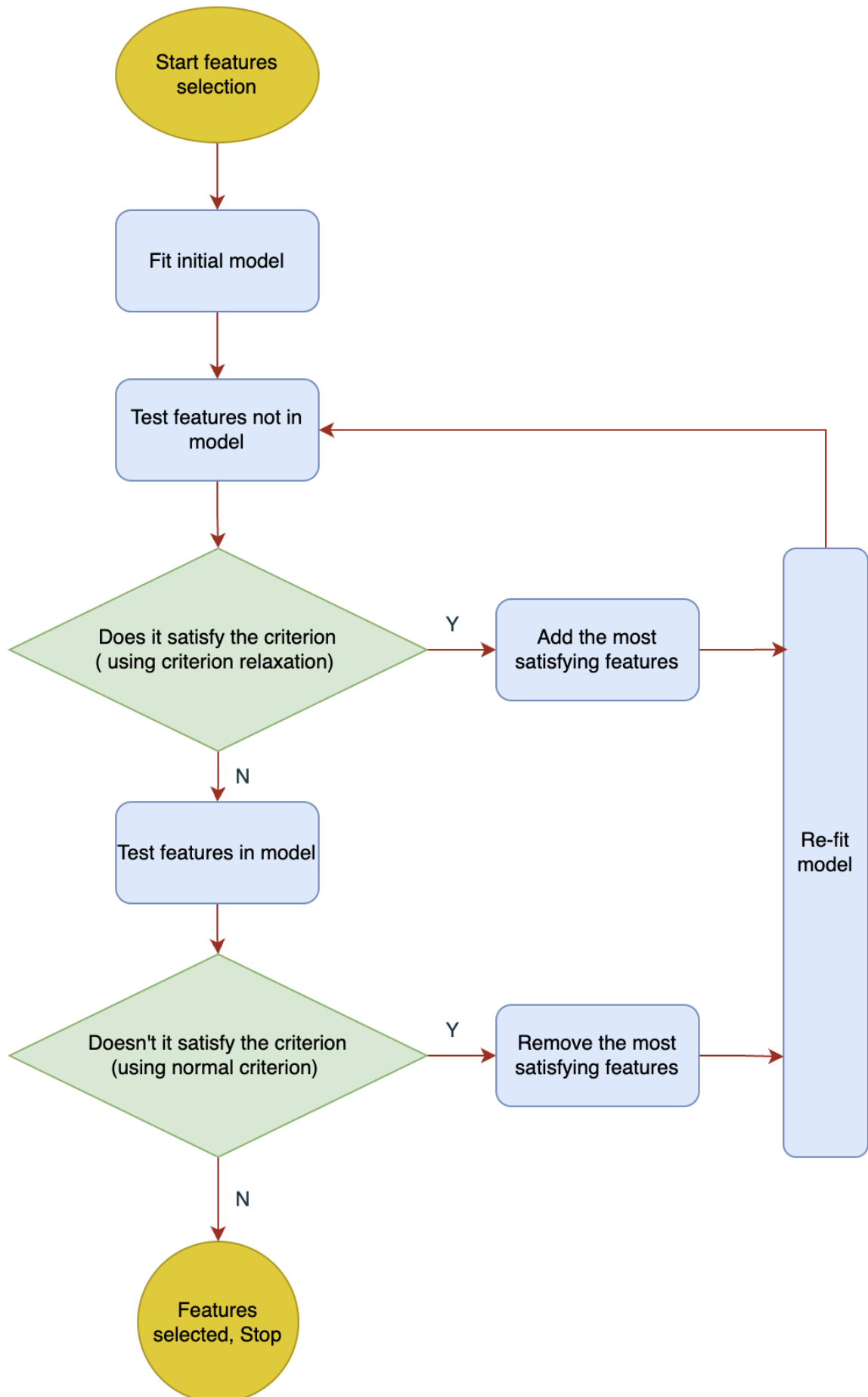
Đầu ra sẽ là các đặc trưng có ý nghĩa (có ảnh hưởng tốt đến kết quả mô hình) được chọn.

MỤC TIÊU

- Cải thiện tốc độ của việc lựa chọn đặc trưng bằng cách đề xuất phương pháp lựa chọn đặc trưng mới
- Đảm bảo hiệu suất tốt của mô hình khi sử dụng tập dữ liệu đầu ra.

NỘI DUNG VÀ PHƯƠNG PHÁP

- ☐ Thu thập các bộ dữ liệu từ các nguồn khác nhau.
- ☐ Áp dụng các phương pháp Feature Selection để chọn ra tập con các đặc trưng quan trọng.
- ☐ Xây dựng các mô hình Machine Learning trên các tập các đặc trưng được chọn.
- ☐ Phân tích một số hạn chế của các phương pháp lựa chọn đặc trưng Forward, Backward, Forward dropping.
- ☐ Thực hiện lựa chọn đặc trưng sử dụng kỹ thuật Criterion relaxation.
- ☐ So sánh các phương pháp dựa trên thực nghiệm để đưa ra nhận xét kết luận.



KẾT QUẢ MONG ĐỢI

- + Hiểu sâu về các kỹ thuật *Feature Selection* và các phương pháp được sử dụng trong *Machine Learning*.
- + Áp dụng *Criterion relaxation* đem lại kết quả tốt hơn về mặt thời gian trong việc lựa chọn đặc trưng và vẫn duy trì hiệu suất tốt.

TÀI LIỆU THAM KHẢO (Định dạng DBLP)

- [1] Stephen A Billings and Hua-Liang Wei. Sparse model identification using a forward orthogonal regression algorithm aided by mutual information. *IEEE Transactions on Neural Networks*, 18(1):306–310, 2007.
- [2] Christophe Couvreur and Yoram Bresler. On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM Journal on Matrix Analysis and Applications*, 21(3):797–808, 2000.
- [3] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse over-complete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.
- [4] A. Aravkin, J. Burke, A. Sholokhov, and P. Zheng, ‘Analysis of Relaxation Methods for Feature Selection in Mixed Effects Models’, *arXiv [stat.ME]*. 2022.
- [5] P. Saha, S. Patikar, and S. Neogy, ‘A Correlation - Sequential Forward Selection Based Feature Selection Method for Healthcare Data Analysis’, in *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*, 2020, pp. 69–72.