Môn học: MÔ HÌNH HÓA THỐNG KÊ
Đồ án kết thúc học phần

Yêu cầu:

- Học viên thực hiện bài làm của mình theo danh sách đã đăng ký, bài làm nào copy lẫn nhau sẽ chia đều số điểm cho các thành viên liên quan.
- Học viên sẽ nộp bài trong ngày 10/08/2024 (0h00 đến 23h59 ngày 10/08/24) trên trang Moodle, không nhận bài nộp trễ (hệ thống sẽ tự đóng sau khi hết hạn nộp). Học viên cần nộp 1 file nén (mỗi nhóm chỉ nộp 1 file nén) gồm 3 file sau:
  - 1 file pdf bài báo cáo (đặt tên file: Nhóm?-DoAn.pdf)
  - 1 file nén các file code (Nhómcode.R hoặc .htlm hoặc .Rmd)
  - 1 file nén gồm các bộ dữ liệu đã sử dụng trong bài.
- Bài báo cáo phải bao gồm tất cả các thông tin sau: Nội dung chính và các phần phụ như mục lục, đề bài, nguồn dữ liệu, hình vẽ, R code, tài liệu tham khảo. Mỗi nhóm cần có bảng phân công công việc rõ ràng cho các thành viên, đánh giá mức độ hoàn thành công việc được giao.
- Bài báo cáo cần được trình bày hợp lý, khoa học.

1. Hoạt động 1 (5 điểm)

1. Dữ liệu được cho trong tập tin "Islander-data.csv" lấy từ

   Htps://www.kaggle.com/datasets/steveahn/memor y-test-on-drugged-islanders-data

   chứa thông tin về một thử nghiệm về tác dụng phụ của các loại thuốc chống trầm cảm đối với trí nhớ của người tham gia thử nghiệm, được đánh giá thông qua thời gian hoàn thành một bài kiểm tra trí nhớ. Người tham gia thử nghiệm sẽ được sử dụng một trong ba loại thuốc khác nhau, với 3 hàm lượng khác nhau và sẽ tiếp xúc với các ký ức vui hoặc buồn trong vòng 10 phút trước khi tiến hành kiểm tra. Thời gian hoàn thành bài kiểm tra của người tham gia sẽ được ghi nhận trước và sau khi kết thúc thử nghiệm để đánh giá hiệu quả của từng loại thuốc cũng như hàm lượng thuốc khác

nhau. (Những người này đều trên 25 tuổi nhằm đảm bảo thuỳ trán phát triển hoàn thiện, nơi đảm nhận chức năng nhận thức và gợi lại ký ức). Dữ liệu được thu thập bởi ông Almohalwas tại UCLA bao gồm 198 quan trắc với 9 biến sau:

- first-name: tên của người tham gia thử nghiệm
- last-name: họ của người tham gia thử nghiệm
- Age: tuổi (năm) của người tham gia thử nghiệm
- HappySadgroup: loại ký ức được tiếp xúc trước khi kiểm tra (H: vui, S: buồn)
- Dosage: Mức độ hàm lượng thuốc sử dụng (1: thấp, 2: trung bình, 3: cao)
- Drug: Loại thuốc sử dụng (A: , Alprazolam, T: Triazolam, S: Placebo)
- Mem-Score-Before: Thời gian (giây) cần để hoàn thành bài kiểm tra trước khi tiếp xúc với thuốc chữa trầm cảm
- Mem-Score-After: Thời gian (giây) cần để hoàn thành bài kiểm tra sau khi tiếp xúc với thuốc chữa trầm cảm
- Diff: Chênh lệch giữa thời gian (giây) hoàn thành bài kiểm tra trước và sau khi sử dụng thuốc.

Yêu cầu:

(a) Tiền xử lý dữ liệu nếu cần;

(b) Trực quan hoá dữ liệu: (Data visualization)

(c) Áp dụng phương pháp phân tích phương sai k nhân tố (ANOVA k nhân tố, *kleq*2) để xem xét những nhân tố nào (trong bộ dữ liệu) có ảnh hưởng đến khả năng gợi nhớ của người sử dụng thuốc chữa trầm cảm.

(d) Đưa ra mô hình được xét, kiểm tra các giả định của mô hình. Rút ra nhận xét, kết luận cần thiết.

(e) Có thể đề xuất những phương pháp cải tiến /phân tích khác có thể cho kết quả tốt hơn nếu có thể.

2. Tập tin **CSM.xlsx** là bộ dữ liệu CSM (Conventional and Social Media Movies) cung cấp một số thuộc tính của phim ảnh lấy từ nguồn *UCI Machine Learning Repository*. Bộ dữ liệu gồm 231 quan trắc trên 14 biến:

   - "Movie": tên phim,
   - "Year": năm phát hành,

- "Ratings": điểm đánh giá,
- "Genre": thể loại phim,
- "Gross": tổng doanh thu,
- "Budget": tổng chi phí,
- "Screens": số rạp chiều,
- "Sequel": phần phim,
- "Sentiment": ý kiến khán giả,
- "Views": số lượt xem,
- "Likes": số lượt thích,
- "Dislikes": số lượt chê,
- "Comments": số bình luận,
- "Aggregate Followers": số người theo dõi.

Yêu cầu:

(a) Tiền xử lý dữ liệu nếu cần;

(b) Chia bộ dữ liệu làm 2 phần: mẫu huấn luyện (training dataset) và mẫu kiểm tra (validation dataset);

(c) Chọn mô hình tốt nhất giải thích cho biến phụ thuộc là biến doanh thu "Gross"thông qua việc chọn lựa các biến độc lập phù hợp trong các biến còn lại từ mẫu huấn luyện. Cần trình bày từng bước phương pháp chọn, tiêu chuẩn chọn mô hình, lý do chọn phương pháp đó.
Kiểm tra các giả định (giả thiết) của mô hình (nếu giả thiết không thoả, có thể biến đổi "transformation"biến (bằng phương pháp Box-Cox, ...), hoặc có thể dùng phương pháp phi tham số, để giải quyết vấn đề này).
Nêu ý nghĩa của mô hình đã chọn.

(d) Dự báo (Prediction): sử dụng mẫu kiểm tra (validation dataset) và dựa vào mô hình tốt nhất được chọn trên đưa số liệu dự báo cho biến phụ thuộc "Gross".

(e) So sánh kết quả dự báo với giá trị thực tế của "Gross". Rút ra nhận xét?

(f) Có thể đề xuất những phương pháp cải tiến /phân tích khác có thể cho kết quả tốt hơn nếu có thể.

2. Hoạt động 2 (5 điểm)

Mỗi nhóm tự chọn 3 bộ dữ liệu (đề tài tự do) để tìm hiểu "nghiên cứu" nhằm vận dụng những kiến thức đã học được từ học phần này để hoàn thành yêu cầu đặt ra.

Học viên có thể tìm dữ liệu từ nhiều nguồn khác nhau, có thể tử hai trang web được cô chọn dữ liệu cho "Hoạt động 1" như trên, ... Lưu ý, bộ dữ liệu được chọn phải có ít nhất 5 biến độc lập và không được lấy lại dữ liệu đã được sửa trên lớp trong các bài tập chọn mô hình, phương pháp ANOVA 2 nhân tố đã làm trên lớp.

Yêu cầu:

- Tên "đề tài", nguồn gốc của dữ liệu, giới thiệu các biến, ...
- Tiền xử lý số liệu,
- Xây dựng mô hình hồi quy tuyến tính bội/ hồi quy trên thành phần chính /hoặc ANOVA 2 hay nhiều nhân tố,
- Kiểm tra các giả định của mô hình hồi quy tuyến tính bội/ hồi quy trên thành phần chính / hoặc ANOVA 2 hay nhiều nhân tố, trên mô hình đã xây dựng,
- Chọn mô hình phù hợp nhất,
- Phân tích kết quả,
- Đưa ra những phương pháp/phân tích khác có thể cho kết quả tốt hơn nếu có thể,
- Kết luận.

Lưu ý: *Đối với học viên thuộc ngành "Giáo dục toán học"học viên bỏ qua mô hình hồi quy trên thành phần chính. Những học viên thuộc hai ngành còn lại, học viên chọn 3 bộ dữ liệu phù hợp cho việc xây dựng tương ứng 3 mô hình như trong yêu cầu (mô hình hồi quy tuyến tính bội, mô hình hồi quy trên thành phần chính và mô hình ANOVA 2 hay nhiều nhân tố).*

Hết.