FINAL PROJECT

Applied Statistics for Engineers and Scientists II

Ngày 29 tháng 7 năm 2024

Yêu cầu:

- Mỗi nhóm sẽ phụ trách một đề tài bao gồm 2 hoạt động. Các nhóm thực hiện riêng bài làm của nhóm mình, bài làm nào copy lẫn nhau sẽ chia đều số điểm cho các thành viên liên quan.
- Mỗi nhóm nộp một báo cáo đề tài dài không quá 60 trang (không tính phụ lục, file code R để trong phần phụ lục). Trên trang bìa của bài báo cáo phải ghi rõ các thông tin sau: Họ tên sinh viên, MSSV, số thứ tự nhóm của mình.
- Sinh viên dùng R/Rstudio để thực hiện việc phân tích dữ liệu trong đồ án này.
- Bài báo cáo phải bao gồm tất cả các thông tin sau: Thu thập và làm sạch dữ liệu (nếu có), mô tả dữ liệu, phân tích thống kê. Ngoài ra, bài báo cáo cần có mục lục, đề bài, hình vẽ, R code, tài liệu tham khảo, nguồn dữ liệu và bảng phân công việc của các thành viên. Bài báo cáo cần trình bày hợp lý, khoa học.
- Bài báo cáo cần nói rõ mục đích ý nghĩa về đề tài của mình, đặc trưng của dữ liệu, phương pháp thống kê, phân tích thống kê và kết quả phân tích, kết luận thống kê, đưa ra đề xuất, thảo luận thêm, ...

Khuyến khích: Sinh viên có thể sử dụng các dữ liệu mới, có ý nghĩa cho chuyên ngành của mình: dữ liệu từ thí nghiệm thực tế, dữ liệu do sinh viên tự thu thập được, ...

Hạn nộp bài báo cáo: 20/8/2024

- Mỗi nhóm cần nộp 1 file nén qua link trên trang Moodle gồm:
  - 1 file pdf bài báo cáo
  - 1 file nén các file code (.R hoặc .htlm hoặc .Rmd)
  - 1 file nén gồm các bộ dữ liệu đã sử dụng trong bài.
- Mỗi nhóm cần nộp 1 bản cứng (bản in) bài báo cáo vào sáng ngày 21/8/2024 (10h00-11h00) tại phòng F206 cho thầy Nguyễn Hữu Toàn.
- Sinh viên cần ký tên vào bảng điểm.

Đề tài

Hoạt động 1: (4 điểm)

Dữ liệu được cho trong file "auto-mpg.csv" là bộ dữ liệu tiêu thụ nhiên liệu của xe trong thành phố. Dữ liệu được lấy từ UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/Auto+MPG

Bộ dữ liệu gồm 398 quan trắc trên 9 biến sau:

- "mpg": (continuous) mức tiêu thụ nhiên liệu tính theo dặm trên galon (miles/gallon),
- "cylinders": (multi-valued discrete) số xy lanh,
- "displacement" : (continuous) kích thước động cơ,
- "horsepower" : (continuous) công suất động cơ,
- "weight" : (continuous) khối lượng,
- "acceleration" : (continuous) gia tốc xe,
- "model year": (multi-valued discrete) năm sản xuất model (2 số cuối)
- "origin": (multi-valued discrete) nơi sản xuất: 1 - North American, 2 - Europe, 3 - Asia
- "car name": (multi-valued discrete) tên xe

Yêu cầu:

1. Nhập và "làm sạch" dữ liệu (lưu ý, biến "horsepower" có 6 quan trắc thiếu dữ liệu; xét xem có dữ liệu ngoại lai không?), thực hiện các thống kê mô tả. (*Chú ý các cột của file "auto-mpg.csv" được phân tách bởi dấu ";", khi đọc file dữ liệu dùng lệnh "read.csv" cần thêm sep = ";"*)
2. Chia bộ dữ liệu làm 2 phần (80/20): mẫu huấn luyện (training dataset) và mẫu kiểm tra (validation dataset).
3. Chọn mô hình tốt nhất giải thích cho biến phụ thuộc "mpg" thông qua việc chọn lựa các biến độc lập phù hợp trong 8 biến độc lập 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.
4. Kiểm tra các giả định (giả thiết) của mô hình.
5. Nêu ý nghĩa của mô hình đã chọn.
6. 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 "mpg". Gọi kết quả dự báo này là biến "predict_mpg".
7. So sánh kết quả dự báo "predict_mpg" với giá trị thực tế của "mpg". Rút ra nhận xét?

Hoạt động 2: (6 điểm)

Mỗi nhóm tự tìm 2 bộ dữ liệu với đề tài tự chọn, khuyến khích sinh viên sử dụng dữ liệu thực tế sẵn có từ các thí nghiệm, khảo sát, dự án, ... trong chuyên ngành của mình. Ngoài ra sinh viên có thể tự tìm kiếm dữ liệu từ những nguồn khác như:

- *https://archive.ics.uci.edu/ml/datasets*
- *https://www.kaggle.com/datasets*
- ...

Dữ liệu cần có ít nhất 1 biến định tính và ít nhất 3 biến định lượng và thực hiện các yêu cầu sau:

- Mô tả dữ liệu, làm sạch dữ liệu nếu cần.
- Thống kê mô tả về dữ liệu. (1 điểm)
- Chọn 1 biến định lượng và thực hiện mô hình hồi quy của biến đó theo ít nhất 4 biến còn lại (có thể sử dụng biến dummy nếu cần trong mô hình hồi quy) nhằm đạt được các mục tiêu của đề tài mình chọn và thực hiện các yêu cầu sau: (2.5 điểm)
  - Phân tích và diễn giải kết quả
  - Nhận xét, rút ra kết luận và đưa ra đề xuất (nếu có) từ kết quả thu được từ mô hình hồi quy trên.