ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC KINH TẾ-LUẬT
KHOA HỆ THỐNG THÔNG TIN
Ngày thi: 09/04/2024

KIỂM TRA GIỮA KỲ
MÔN: **Phân tích dữ liệu với R/Python**
Lớp: **K21411C**
(Đề thi gồm 2 trang)
Thời gian làm bài: **120** phút

***Quy định*:**

- ✓ *Được sử dụng mọi tài liệu thuộc sở hữu của mỗi SV.*
- ✓ *Sau khi hết giờ làm bài, SV tự nén thư mục lưu kết quả bài thi theo quy tắc đặt tên* ***MSSV_HoTenSV.rar/zip*** *và nộp bài trên hệ thống LMS theo hướng dẫn của giảng viên coi thi.*

Đề bài

Câu 1: (1đ) Theo Anh/Chị liên quan đến lĩnh vực khoa học dữ liệu có những ngành nghề nào? Để đáp ứng được yêu cầu công việc với những ngành nghề đó, Anh/Chị cần chuẩn bị những gì?

Câu 2: (1đ) Với dữ liệu về *giá xe ($1000)*, *số năm sử dụng*, *số km đã đi (1000Km)* được cho trước ở tập tin **XeDaQuaSuDung.xml**. Hãy thực hiện các yêu cầu sau:

a) Vẽ biểu đồ phân tán thể hiện tương quan giá xe theo số năm sử dụng, giá xe theo số km đã đi? Nhận xét?

b) Vẽ biểu đồ nhiệt (heatmap) thể hiện mức tương quan giữa số năm sử dụng, số km đã đi đối với giá xe.

c) Ước lượng các hệ số của mô hình hồi quy sau:

$$\widehat{Giá\ xe} = a + b*(số\ năm\ sử\ dụng) + c*(số\ km\ đã\ đi)$$

d) Phát biểu ý nghĩa của các hệ số hồi quy ước lượng?

e) Kiểm định độ phù hợp của mô hình với mức ý nghĩa 5%, diễn giải ý nghĩa hệ số xác định?

f) Lưu mô hình và thực hiện dự đoán giá xe với các thông tin sau:

- Xe sử dụng 6 năm, đi được 112000Km?
- Xe sử dụng 3 năm, đi được 165000Km?

Câu 3: (2đ) Với dữ liệu ở tập tin **Spending_data.json**. Hãy thực hiện các yêu cầu sau:

a) Vẽ biểu đồ nhiệt (heatmap) thể hiện tương quan giữa các biến? Nhận xét các yếu tố ảnh hưởng đến "Expenditure (VND)"?

b) Xây dựng mô hình hồi quy và phát biểu ý nghĩa của các hệ số hồi quy ước lượng? Kiểm định độ phù hợp của mô hình với mức ý nghĩa 5%?

c) Thực hiện đánh giá, lưu mô hình và dự đoán chi tiêu (expenditure) trong tương lai.

Câu 4: (2.5đ) Thu thập dữ liệu cổ phiếu VNM từ năm 2013 đến năm 2023 lưu với tập tin **VNM_2013_2023.xlsx**, hãy thực hiện những yêu cầu sau:

a) Phân rã dữ liệu? Nhận xét?

b) Kiểm định tính dừng, tương quan.

c) Fit mô hình ARIMA với 80% dữ liệu huấn luyện, đánh giá mô hình và dự báo giá cổ phiếu trong tương lai.

d) Thực hiện chia dữ liệu theo kỹ thuật Expand Window, đánh giá mô hình.

e) Thực hiện các yêu cầu tương tự như trên với dữ liệu tập tin **Data.xls**

Câu 5: (2.5đ) Với dữ liệu về doanh thu (tỷ đồng) và chi phí đầu tư quảng cáo (triệu đồng) qua các kênh như youtube, facebook, newspaper được cung cấp ở tập tin **Ads.txt**. Hãy thực hiện các yêu cầu sau:

a) Mô tả dữ liệu với các trị thống kê: min, max, tứ phân vị.

b) Vẽ biểu đồ pairplot và heatmap thể hiện mức tương quan giữa chi phí đầu tư qua các kênh với doanh thu. Cho biết giá trị tương quan cụ thể?

c) Sử dụng mô hình cây quyết định tiên lượng doanh thu qua chi phí đầu tư quảng cáo với doanh thu được phân hoạch thành 3 mức sau:

- Cho biết số lượng mẫu tin theo doanh thu ứng với 3 mức phân hoạch?
- Xây dựng mô mình với 80% số mẫu dữ liệu ngẫu nhiên.
- Kiểm thử mô hình với 20% số mẫu dữ liệu còn lại. Cho biết kết quả kiểm thử qua các độ đo: Precision, Recall, F1-Score. Đánh giá kết quả dự đoán qua ma trận hỗn loạn (confusion matrix).
- Thực hiện chia dữ liệu k-fold và đánh giá chéo (cross-validation)

d) Trực quan hóa cây quyết định.

e) Dự báo mức doanh thu với chi phí quảng cáo (triệu đồng) như sau:

- youtube: 120, facebook: 65, newspaper: 20
- youtube: 35, facebook: 45, newspaper: 15

Câu 6: (1đ) Để đảm bảo mô hình có độ chính xác và khả năng tổng quát tốt, Anh/Chị đã thực hiện kỹ thuật phân chia dữ liệu theo k-fold và cross-validation như thế nào? Anh/Chị xử lý vấn đề dữ liệu overfitting hoặc underfitting khi xây dựng mô hình như thế nào?

**Lưu ý: Đối với các câu từ 2 đến 5, SV có thể sử dụng ngôn ngữ Python hoặc R để thực hiện. Thư mục bài nộp bao gồm mã nguồn chương trình, dữ liệu phân tích và tập tin word diễn giải chi tiết kết quả phân tích.*

----------HẾT----------