TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

ĐỀ THI KẾT THÚC HỌC PHẦN
Học kỳ 1 – Năm học 2024-2025

MÃ LƯU TRỮ
(do phòng KT-ĐBCL ghi)
CK2425.1
CSC14119

Tên học phần: Nhập môn Khoa học Dữ liệu **Mã HP:** CSC14119

Thời gian làm bài: **100 phút** Ngày thi: **09/01/2025**

Ghi chú: *Sinh viên* [☒ *được phép* / ☐ *không được phép*] *sử dụng tài liệu khi làm bài.*

Họ tên sinh viên: ..... **MSSV**: ..... **STT**: .....

Câu 1 (*1.5 điểm*).

a. Một khảo sát trong năm 2016 chỉ ra rằng 79% thời gian của các dự án Khoa học Dữ liệu được dành cho việc chuẩn bị dữ liệu. Hãy phân tích các yếu tố ảnh hưởng đến thời gian chuẩn bị dữ liệu và giải thích tại sao việc chuẩn bị dữ liệu lại chiếm một phần lớn như vậy trong quy trình phân tích dữ liệu.

b. Một trong những nguyên tắc quan trọng khi thu thập dữ liệu từ các trang web là tôn trọng quyền riêng tư và tuân thủ quy định của trang web. Hãy giải thích cách kiểm tra các giới hạn thu thập dữ liệu của một trang web bằng nhiều cách khác nhau.

Câu 2 (3 *điểm*).

Dữ liệu về mức thu nhập hàng tháng (triệu đồng) của một nhóm 50 người được chọn ngẫu nhiên trong trường ĐH.Khoa học Tự nhiên như sau:

5, 6, 7.5, 8, 4, 6.5, 12, 5.5, 9, 11, 0.5, 7, 6.2, 7.2, 6.7, 1, 7, 8.5, 9.5, 10, 6.8, 177.1, 7, 8.1, 6.2, 7.5, 7, 49.8, 4, 4.3, 8.

a. Tính trung bình (mean), trung vị (median), yếu vị (mode), phương sai (variance) và độ lệch chuẩn (standard deviation) của dữ liệu.

b. Tính tứ phân vị (Q1, Q3) và khoảng tứ phân vị (IQR).

c. Xác định các giá trị ngoại lai (outliers) dựa trên tứ phân vi.

d. Dựa trên dữ liệu, bạn có nhận xét gì về phân phối thu nhập (ví dụ: tập trung, phân tán, bất đối xứng)?

e. Các giá trị ngoại lai có ảnh hưởng như thế nào đến trung bình? Nếu loại bỏ các giá trị ngoại lai, hãy tính lại trung bình.

f. Vẽ biểu đồ box plot để thể hiện các thông tin thống kê cơ bản và chỉ ra các giá trị ngoại lai (nếu có).

Câu 3 (3 *điểm*).

Bạn được giao nhiệm vụ phân tích dữ liệu khám chữa bệnh của bệnh viện XHCMUS. Dữ liệu đến từ 2 bộ phận khác nhau trong bệnh viện như sau:

Bảng 1. Dữ liệu khám sức khỏe

Mã khách hàng	Tên	Ngày khám	Tuổi	BMI	Huyết áp (mmHg)	Tình trạng
KH001	Nguyễn Văn A	12/1/2023	35	23.5	120/80	Tốt
KH002	Trần Thị B	2023-2-25	29		130/85	Khá
KH003	Phạm Văn C	12/5/2023	40	27.8		Yếu
KH004	Lê Thị D	12/2/2023	38		140/90	
KH005	Nguyễn Thị E	12/4/2023	50	22.1	125/82	Tốt

Bảng 2. Dữ liệu đăng ký khách hàng

Mã KH	Họ tên	Ngày sinh	Chiều cao (m)	Cân nặng (kg)	Giới tính	Tỉnh
KH001	Nguyễn Văn A	1/1/1988	1.75	72	Nam	TP. Hồ Chí Minh
KH002	Trần Thị B		1.65	68	Nữ	Hà Nội
KH003	Phạm Văn C	8/15/1983	1.7	90	Nam	Đà Nẵng
KH004	Lê Thị D	Ngày 2 tháng 10 năm 1985	1.7	85		Cần Thơ
KH005	Nguyễn Thị E	9/25/1973	1.6	-57	Nữ	HCM City

a. Hãy tổng hợp 2 bảng trên lại thành một bảng để phục vụ cho quá trình phân tích. Áp dụng các kỹ thuật khác nhau cùng lý giải tại sao cần áp dụng để làm cho bảng tổng hợp ở trạng thái tốt nhất. Khuyến khích mỗi quá trình như làm sạch dữ liệu (data cleaning), biến đổi dữ liệu (data transformation), giảm dữ liệu (data reduction) đều có đại diện để xử lý.
b. Chọn biểu đồ thích hợp để minh họa mối quan hệ giữa BMI và Tuổi.

Câu 4 (2.5 *điểm*).

Bạn được cung cấp một tập dữ liệu về mối quan hệ giữa số giờ học của sinh viên, số lượng bài tập hoàn thành và điểm số của họ trong kỳ thi. Mục tiêu của bạn là xây dựng mô hình hồi quy tuyến tính dự đoán điểm số dựa trên số giờ học và số bài tập hoàn thành.

Số giờ học	Số bài tập hoàn thành	Điểm số (y)
1.5	3	5.0
2	5	5.5
2.5	6	6.0
3	8	6.5
3.5	9	7.0
4	10	7.5
4.5	12	8.0

a. Sử dụng phương pháp Min-Max Normalization để chuẩn hóa cột Số giờ học và Số bài tập hoàn thành về khoảng [0, 1] trước khi xây dựng mô hình.
b. Giả sử dữ liệu tuân theo hàm hồi quy tuyến tính, hãy sử dụng hàm lỗi MSE và phương pháp gradient descent qua 2 vòng lặp với tỉ lệ học $\eta = 0.01$ để tìm ra biểu diễn của hàm này. Các tham số đều được khởi tạo là 0.

– Hết –