TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

ĐỀ THI KẾT THÚC HỌC PHẦN

Học kỳ 1–Năm học 2023-2024 (CTĐA-CNTT)

MÃ LƯU TRỮ
(do phòng KT-ĐBCL ghi)
DA_CK2324.1
CSC14004

Tên học phần: Khai thác dữ liệu và ứng dụng **Mã HP:** CSC14004

Thời gian làm bài: **90 phút** Ngày thi: **27/12/2023**

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

(Không sử dụng Laptop & Smart Phone)

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

Câu 1: Xét tập dữ liệu các tai nạn giao thông trong bảng sau:

Weather Condition	Driver's Condition	Traffic Violation	Seat Belt	Crash Severity
Good	Alcohol-impaired	Exceed speed limit	No	Major
Bad	Sober	None	Yes	Minor
Good	Sober	Disobey stop sign	Yes	Minor
Good	Sober	Exceed speed limit	Yes	Major
Bad	Sober	Disobey traffic signal	No	Major
Good	Alcohol-impaired	Disobey stop sign	Yes	Minor
Bad	Alcohol-impaired	None	Yes	Major
Good	Sober	Disobey traffic signal	Yes	Major
Good	Alcohol-impaired	None	No	Major
Bad	Sober	Disobey traffic signal	No	Major
Good	Alcohol-impaired	Exceed speed limit	Yes	Major
Bad	Sober	Disobey stop sign	Yes	Minor

a. Hãy biểu diễn lại tập dữ liệu dưới dạng nhị phân.

b. Chiều dài lớn nhất của mỗi giao dịch trong dữ liệu ở câu a là bao nhiêu?.

c. Giả sử rằng *minSup* = 30%, có bao nhiêu Itemsets ứng viên và phổ biến sẽ được phát sinh?

d. Tạo một tập dữ liệu chỉ chứa các thuộc tính nhị phân bất đối xứng sau: **(Weather = Bad, Driver's condition = Alcohol-impaired,Traffic violation = Yes, Seat Belt = No, Crash Severity = Major)**. Với **Traffic violation** chỉ None thì có giá trị là 0, các giá trị còn lại thì gán bằng 1. Giả sử minSup = 30%, cho biết có bao nhiêu Itemsets ứng viên và phổ biến được phát sinh?

e. So sánh số lượng Itemsets ứng viên và phổ biến trong câu c và d

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

ĐỀ THI KẾT THÚC HỌC PHẦN

Học kỳ 1–Năm học 2023-2024 (CTĐA-CNTT)

MÃ LƯU TRỮ
(do phòng KT-ĐBCL ghi)

Câu 2: Cho CSDL sản phẩm DB_e như sau:

Loại sản phẩm	*PID*	*Items*	*Val* (nghìn đô-la)
P_1	1	*a, b, c*	2100
P_2	2	*a, c*	1000
P_3	3	*a, b*	1000
P_4	4	*b, c, d*	150
P_5	5	*c, d*	50
P_6	6	*b, d*	100
P_7	7	*c, d, e, f, g*	200
P_8	8	*d, e, f, h*	100
P_9	9	*e, f*	50
P_{10}	10	*b, e, h*	100
P_{11}	11	*e, c*	150

Cơ sở dữ liệu sản phẩm DB_e gồm 11 loại sản phẩm và tập toàn bộ các thành phần cấu tạo nên các loại sản phẩm gồm 8 thành phần là $\{a, b, c, d, e, f, g, h\}$. Tổng lợi nhuận của các loại sản phẩm là 5000 nghìn đô-la. Giả sử ngưỡng giảm lợi nhuận đặt ra là $\xi = 18\%$. Sử dụng thuật toán META tìm tất cả các tập thành phần không hữu ích trong DB_e.

Câu 3: Cho bảng quan sát về thời tiết như sau:

Ngày	Quang cảnh	Nhiệt độ	Độ ẩm	Gió	Chơi Tennis?
1	Nắng	Nóng	Cao	Thấp	Không đi
2	Âm u	Nóng	Cao	Thấp	Đi
3	Mưa	Lạnh	TB	Cao	Không đi
4	Âm u	TB	Cao	Thấp	Đi
5	Mưa	TB	Cao	Thấp	Đi
6	Mưa	Lạnh	TB	Thấp	Đi
7	Nắng	TB	Cao	Thấp	Không đi
8	Nắng	Lạnh	TB	Thấp	Đi
9	Âm u	Lạnh	TB	Cao	Đi
10	Mưa	TB	TB	Thấp	Đi
11	Nắng	Nóng	Cao	Cao	Không đi
12	Nắng	TB	TB	Cao	Đi
13	Âm u	TB	Cao	Cao	Đi
14	Âm u	Nóng	TB	Thấp	Đi
15	Mưa	TB	Cao	Cao	Không đi

Cho biết kết quả của mẫu X= (Nắng, Nóng, TB, Cao):

a. Theo phương pháp k-NN với k = 3.

b. Theo phương pháp Naïve Bayesian.

HẾT