

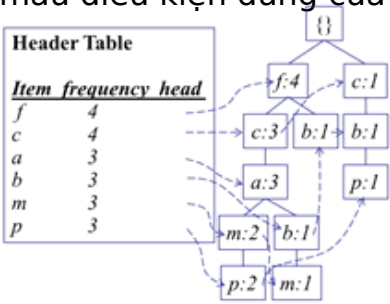
Câu 1: Khai thác dữ liệu có thể giúp cho doanh nghiệp?	Đáp án đúng: Cải thiện chiến lược kinh doanh.
Câu 2: Khai thác dữ liệu là bước phát triển tiếp theo của?	Đáp án đúng: Khoa học về dữ liệu
Câu 3: Tri thức khám phá được từ quy trình khai thác dữ liệu được sử dụng bởi ai?	Đáp án đúng: Chuyên viên hoạch định chiến lược
Câu 4: Dữ liệu lịch sử (historical data) được hiểu là	Đáp án đúng: Dữ liệu hình thành trong toàn bộ quá trình kinh doanh.
Câu 5: Vấn đề nào sau đây KHÔNG phải là khó khăn (thách thức) trong khai phá dữ liệu?	Đáp án đúng: Có quá nhiều thuật toán
Câu 6: Phân tích giỏ hàng là bài toán	Đáp án đúng: Tìm luật kết hợp
Câu 7: Hệ thống nào sau đây sử dụng khai thác dữ liệu ?	Đáp án đúng: Hệ thống hỗ trợ quyết định
Câu 8: Lọc thư rác là ứng dụng của bài toán	Đáp án đúng: Phân lớp
Câu 9: Một siêu thị muốn sắp xếp vị trí của các món hàng trên kệ hàng, sao cho có thể tối đa hóa lợi nhuận. Vậy, siêu thị nên thực hiện bài toán gì trên lịch sử mua hàng của khách hàng?	Đáp án đúng: Phân tích luật kết hợp
Câu 10: Giai đoạn nào được thực hiện đầu tiên trong quy trình phát hiện tri thức?	Đáp án đúng: Làm sạch dữ liệu
Câu 11: Hãy chọn phát biểu ĐÚNG trong các phát biểu sau:	Đáp án đúng: Khai thác dữ liệu cho dù được sử dụng bởi tổ chức nào cũng gây nên tác động xã hội.
Câu 12: Tích hợp dữ liệu nhằm	Đáp án đúng: Tập hợp dữ liệu từ nhiều nguồn thành một khối.
Câu 13: Thao tác nhóm các mẫu dữ liệu gần giống nhau vào những nhóm chưa xác định trước. Chúng ta sử dụng kỹ thuật	Đáp án đúng: Học không giám sát

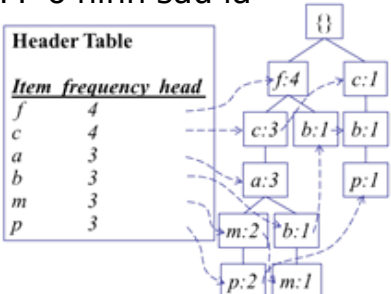
Câu 14: Một trong những kỹ thuật để phát hiện ngoại lệ (outlier detection) là sử dụng bài toán	Đáp án đúng: Khai thác luật kết hợp
Câu 15: Một công ty viễn thông muốn nhóm những khách hàng của họ vào những nhóm riêng biệt để triển khai chương trình khuyến mãi phù hợp cho từng nhóm. Đây là ví dụ của	Đáp án đúng: Học không giám sát
Câu 16: Cho bảng dữ liệu sau về kế hoạch vào đại học của các em học sinh Học sinh Điểm thi ĐH Gia đình khuyến khích Học bổng (ngàn đồng/tháng) Kế hoạch vào ĐH An 21 Có 800 Có Hòa 23 Không 1000 Có Thái 18 Có 0 Có Bình 28 Không ? Không Giả sử bạn là một mệnh thường quân muốn xác định mức học bổng hàng tháng để giúp các học sinh giỏi có hoàn cảnh khó khăn được học đại học. Bạn sử dụng bài toán gì để xác định số tiền học bổng cho em Bình để em thay đổi kế hoạch vào ĐH từ “không” thành “có”?	Đáp án đúng: Hồi quy
Câu 17: Hệ thống khuyến nghị (recommender systems) trong tiếp thị chéo (cross-marketing) thường sử dụng kết quả của	Đáp án đúng: Luật kết hợp
Câu 18: Khi khai thác dữ liệu được sử dụng phổ biến, phát biểu nào sau đây là KHÔNG ĐÚNG	Đáp án đúng: Người dùng có thể kiểm soát được thông tin nào của mình có thể được khai thác.
Câu 19: Để tư vấn ngành học cho thí sinh A muốn xét tuyển vào đại học, dựa vào thông tin cá nhân của thí sinh A và một lượng lớn sinh viên đang theo học hoặc đã ra trường, ta cần làm bài toán:	Đáp án đúng: Phân lớp
Câu 20: Câu hỏi “Khách hàng có thông tin là X có tiềm năng hay không?” có thể trả lời bằng:	Đáp án đúng: Phân lớp
Câu 21: Để xác định một khách hàng tiềm năng ta sử dụng bài toán	Đáp án đúng: Phân lớp

Câu 22: Câu hỏi “Một khách hàng là giáo viên thường mua gì sau khi mua máy tính?” được trả lời bởi bài toán nào?	Đáp án đúng: Khai thác luật kết hợp đa chiều
Câu 23: Thói quen mua hàng của những nhóm khách hàng khác nhau trong siêu thị có thể được phát hiện nhờ phương pháp	Đáp án đúng: Phân cụm
Câu 24: Cho một cơ sở dữ liệu về lý lịch, điểm số về quá trình học tập của sinh viên. Bài toán phân lớp có thể:	Đáp án đúng: Dự đoán xếp loại tốt nghiệp của một sinh viên.
Câu 25: Với một công ty bán lẻ có khoảng một triệu khách hàng, công ty muốn thực hiện chương trình khuyến mãi sao cho phù hợp với nhu cầu của khách hàng, bước đầu tiên công ty nên làm là	Đáp án đúng: Phân cụm khách hàng dựa vào thông tin cá nhân và hành vi mua hàng.
Câu 26: Thao tác nào sau đây không thuộc giai đoạn tiền xử lý dữ liệu?	Đáp án đúng: Trực quan hóa dữ liệu
Câu 27: Kích cỡ quần áo (size = {XS, S, M, L, XL, XXL,...}) là kiểu dữ liệu	Đáp án đúng: Thứ bậc
Câu 28: Vẽ biểu đồ cho dữ liệu ta gọi là	Đáp án đúng: Trực quan hóa dữ liệu
Câu 29: _____ không phải là một yêu cầu của chất lượng dữ liệu?	Đáp án đúng: Dung lượng của dữ liệu
Câu 30: Cho một tập dữ liệu về độ tuổi bao gồm 10 phần tử như sau: Age = {15, 16, 18, 22, 24, 24, 24, 25, 26, 30} Trung vị của tập dữ liệu Age là?	Đáp án đúng: 24
Câu 31: Cho một tập dữ liệu về độ tuổi bao gồm 10 phần tử như sau: Age = {15, 16, 18, 22, 24, 24, 24, 25, 26, 30} Độ lệch chuẩn của tập dữ liệu Age là?	Đáp án đúng: 4.71
Câu 32: Histogram của phân bố chuẩn có dạng	Đáp án đúng: Hình chuông
Câu 33: Việc thay thế số tuổi bằng giá trị đại diện như thiếu nhi, thiếu niên, thanh niên, trung niên, cao niên... gọi là	Đáp án đúng: Rời rạc hóa dữ liệu

Câu 34: Dữ liệu rời rạc...	Đáp án đúng: Chỉ chấp nhận một số giá trị hữu hạn
Câu 35: Hàm cosine để đo độ tương đồng giữa	Đáp án đúng: Hai vector kiểu số
Câu 36: $d(i, j) = \sqrt[2]{ x_{i1} - x_{j1} ^2 + x_{i2} - x_{j2} ^2 + \dots + x_{ip} - x_{jp} ^2}$ được gọi là công thức tính khoảng cách Euclidean khi h bằng	Đáp án đúng: 2
Câu 37: $\text{sim}(A, B) = \frac{A \cdot B}{\ A\ \ B\ }$ là công thức tính	Đáp án đúng: Độ tương đồng cosine
Câu 38: Hãy cho biết các chỉ số thể hiện trong biểu đồ boxplot là gì?	Đáp án đúng: Min, Max, Q1, Q3, Median
Câu 39: OLAP nghĩa là	Đáp án đúng: Phân tích dữ liệu trực tuyến
Câu 40: Đâu là chức năng chính của OLAP?	Đáp án đúng: Hỗ trợ quyết định
Câu 41: _____ là quá trình lấy dữ liệu từ nhiều nguồn và ánh xạ từng trường vào một cấu trúc trong nhà kho dữ liệu (data warehouse).	Đáp án đúng: Tích hợp dữ liệu
Câu 42: Kiểu Data Mart có thể lấy dữ liệu từ kho dữ liệu hoặc hệ thống vận hành.	Đáp án đúng: Kiểu Data Mart phụ thuộc
Câu 43: Việc tạo cơ sở dữ liệu vật lý và các cấu trúc logic nằm trong giao đoạn nào	Đáp án đúng: Xây dựng
Câu 44: _____ là một tập hợp dữ liệu hướng chủ đề, toàn vẹn, không bị rò rỉ mất mát và có giá trị lịch sử phục vụ cho công tác quản lý, cung cấp thông tin một cách kịp thời, chính xác, đồng thời là nền tảng cho việc xây dựng các ứng dụng phân tích dữ liệu, hỗ trợ ra quyết định của tổ chức.	Đáp án đúng: Data warehouse - Kho dữ liệu
Câu 45: OLTP là tên viết tắt của kĩ thuật nào?	Đáp án đúng: Online Transaction Processing

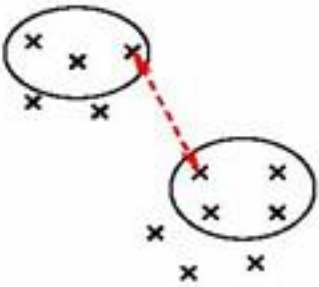
Câu 46: Trong khai phá dữ liệu, thao tác cắt khối (dice) giúp người sử dụng phân tích dữ liệu trên mấy chiều?	Đáp án đúng: Hai chiều trở lên.
Câu 47: _____ dự đoán xu hướng và hành vi trong tương lai, hỗ trợ các nhà quản lý đưa ra quyết định tác động đến hoạt động kinh doanh của doanh nghiệp.	Đáp án đúng: Data mining - Khai phá dữ liệu
Câu 48: Đối với thuật toán Apriori, để tìm tất cả các luật kết hợp, ta cần phải làm gì trước đó?	Đáp án đúng: Tìm tất cả các tập phổ biến
Câu 49: Duyệt cơ sở dữ liệu nhiều lần là yếu điểm của thuật toán	Đáp án đúng: Apriori
Câu 50: Bài toán khai thác tập phổ biến và luật kết hợp KHÔNG được áp dụng cho câu hỏi nào sau đây?	Đáp án đúng: Giá cổ phiếu X lên hay giảm trong 5 tháng tới?
Câu 51: Ứng dụng nào sau đây KHÔNG áp dụng bài toán khai thác tập phổ biến?	Đáp án đúng: Xác định hạn mức tín dụng
Câu 52: Thuật toán FP-Growth cần quét cơ sở dữ liệu bao nhiêu lần?	Đáp án đúng: 2 lần
Câu 53: Việc sắp xếp vị trí các món hàng trong một catalogue có thể được cải tiến nhờ vào bài toán ?	Đáp án đúng: Khai thác luật kết hợp
Câu 54: Theo cơ sở dữ liệu giao tác trong bảng sau, độ tin cậy của luật kết hợp A→B là: TID Itemset 100 A B C D 200 A C D 300 A B D 400 B C D E 500 A B C E	Đáp án đúng: 75%
Câu 55: Theo cơ sở dữ liệu giao tác trong bảng sau, độ hỗ trợ và độ tin cậy của luật kết hợp AC → D lần lượt là: TID Itemset 100 A B C D 200 A C D 300 A B D 400 B C D E 500 A B C E	Đáp án đúng: 40%; 66.7%
Câu 56: Chọn công thức đúng để tính độ tin cậy của luật $A \rightarrow B$	Đáp án đúng: $confidence(A \rightarrow B) = \frac{P(A \cup B)}{P(A)}$

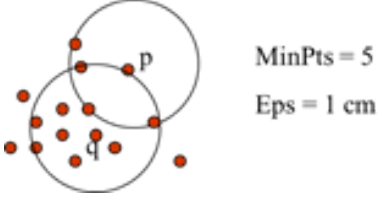
Câu 57: Hãy cho biết trong các đặc điểm sau, đặc điểm nào KHÔNG phải của thuật toán Apriori?	Đáp án đúng: Cần xây dựng cây.
Câu 58: Độ hỗ trợ (support) của một luật kết hợp là	Đáp án đúng: Mức độ phổ biến của luật
Câu 59: Cho s là độ hỗ trợ, c là độ tin cậy của một luật kết hợp. Hãy chọn phát biểu ĐÚNG	Đáp án đúng: $s(A \Rightarrow B) = s(B \Rightarrow A)$ và $c(A \Rightarrow B) \neq c(B \Rightarrow A)$
Câu 60: Hãy chọn phát biểu ĐÚNG trong các phát biểu sau đây?	Đáp án đúng: Độ hỗ trợ tối thiểu (min support) càng cao thì số tập phổ biến được phát hiện càng ít.
Câu 61: Câu nào sau đây KHÔNG đúng về Apriori khi so sánh Apriori và FP-Growth?	Đáp án đúng: Apriori gặp khó khăn khi duyệt cây đệ quy.
Câu 62: Một trong những cách để hạn chế yếu điểm của thuật toán Apriori là?	Đáp án đúng: Giảm số lần quét cơ sở dữ liệu
Câu 63: Bước đầu tiên của thuật toán FP-Growth là?	Đáp án đúng: Sắp xếp các mục trong giao tác theo thứ tự giảm dần của độ hỗ trợ.
Câu 64: Trong hình sau, hãy lựa chọn tập cơ sở mẫu điều kiện đúng của mục 	Đáp án đúng: $\{fca:1\}$, $\{f:1\}$, $\{c:1\}$
Câu 65: Một luật có dạng: $\text{age}(X, "19-25") \wedge \text{occupation}(X, "student") \rightarrow \text{buys}(X, "coke")$ được gọi là luật gì?	Đáp án đúng: Luật kết hợp đa chiều
Câu 66: Để rút trích số lượng tập phổ biến mà không mất mát thông tin, thay vì tìm tất cả các tập phổ biến, ta đi tìm	Đáp án đúng: Tất cả các tập phổ biến đóng (closed patterns).

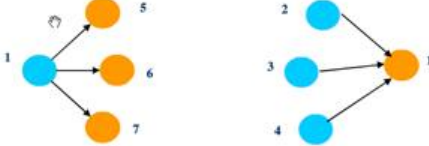
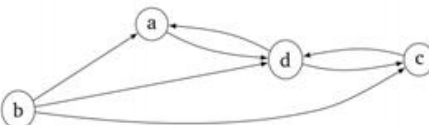
Câu 67: Cho X là một tập mục, X là tập phổ biến đóng (closed-pattern) khi và chỉ khi X là phổ biến và	Đáp án đúng: Không tồn tại một tập mục Y sao cho $Y \supset X$, mà Y có cùng độ hỗ trợ với X.
Câu 68: Thực thi thuật toán Apriori, nếu kết quả cuối cùng ta tìm được các tập phổ biến có độ dài bằng n, ta phải duyệt CSDL tối đa bao nhiêu lần?	Đáp án đúng: $n+1$
Câu 69: Áp dụng Apriori, với minsup=3, nếu lần lặp thứ hai ta đếm độ hỗ trợ của các tập mục được {AB:3}, {AC:2}, {BC:3}, {BD:4}. Vậy tập ứng viên ở lần lặp thứ 3 là?	Đáp án đúng: $C_3 = \{ABD, BCD\}$
Câu 70: Các tập phổ biến chứa m khi duyệt cây FP ở hình sau là 	Đáp án đúng: {m:3}, {fm:3}, {cm:3}, {fcm:3}, {fam:3}, {fcam:3}
Câu 71: Từ nào sau đây KHÔNG liên quan đến thuật toán cây quyết định?	Đáp án đúng: Quy hoạch động
Câu 72: Câu nói “Hãy cho tôi biết bạn chơi với ai, tôi sẽ nói cho bạn biết bạn như thế nào” có thể dùng khi nói về thuật toán nào?	Đáp án đúng: KNN
Câu 73: Thuật toán nào sau đây tốn thời gian tính khoảng cách giữa các đối tượng dữ liệu?	Đáp án đúng: KNN
Câu 74: Naïve Bayes là thuật toán phân lớp dựa trên gì?	Đáp án đúng: Thống kê
Câu 75: SVM là viết tắt của	Đáp án đúng: Support Vector Machine

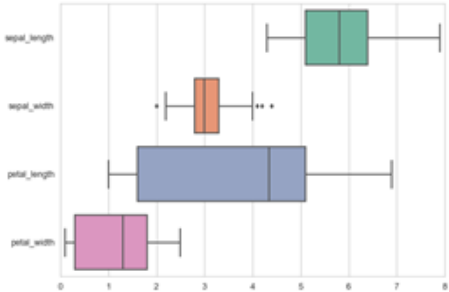
Câu 76: Trong bài toán dự đoán độ tin cậy của khách hàng vay tiền ngân hàng. Trường hợp khách hàng là đáng tin cậy nhưng kết quả dự đoán là không tin cậy. Vậy kết quả dự đoán được gọi là	Đáp án đúng: False Negative
Câu 77: Dựa vào bảng sau, giá trị của xác suất $P(\text{Nghề nghiệp} = \text{"CNV"} \text{An toàn} = \text{"Có"})$ là: Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có	Đáp án đúng: 2/5
Câu 78: Thuật toán phân lớp dựa vào láng giềng gần nhất thích hợp nhất để xử lý dữ liệu kiểu gì?	Đáp án đúng: Số liên tục
Câu 79: Thuật toán nào sau đây sử dụng Entropy làm độ đo thông tin?	Đáp án đúng: ID3
Câu 80: Trong các độ đo để lựa chọn thuộc tính rẽ nhánh khi dựng cây quyết định, độ đo nào dễ bị nghiêng về thuộc tính có nhiều giá trị?	Đáp án đúng: Information Gain
Câu 81: Nếu thuật toán dựng cây quyết định bị overfitting thì điều gì sau đây không xảy ra?	Đáp án đúng: Thuật toán dựng cây bị rơi vào vòng lặp vô hạn
Câu 82: Thuật toán phân lớp nào sau đây chỉ thực hiện trên dữ liệu kiểu số	Đáp án đúng: SVM
Câu 83: Độ bao phủ (recall) của phép tiên đoán trong bảng sau là Actual\Prediction Cancer Not Cancer Total Cancer 1300 1200 2500 Not Cancer 2700 4800 7500 Total 4000 6000 10000	Đáp án đúng: 52%
Câu 84: Độ tin cậy (precision) của phép tiên đoán trong bảng sau là Actual\Prediction Cancer Not Cancer Total Cancer 1300 1200 2500 Not Cancer 2700 4800 7500 Total 4000 6000 10000	Đáp án đúng: 32.5%

Câu 85: Thuật toán Naïve Bayes có thể cho độ chính xác không cao do điều gì?	Đáp án đúng: Cân giả định độc lập về mặt thống kê của các thuộc tính.
Câu 86: Công thức sau được sử dụng trong thuật toán nào? $Info(D) = -\sum_{i=1}^M p_i \log_2(p_i)$	Đáp án đúng: ID3
Câu 87: Một bệnh nhân không tin tưởng kết quả chẩn đoán của bác sỹ. Bệnh nhân đó quyết định đi khám lần lượt nhiều bác sỹ và chọn kết quả được nhiều bác sỹ chọn nhất. Đây là ý tưởng của phương pháp tập hợp mô hình nào?	Đáp án đúng: Bagging
Câu 88: Trong phép phân lớp, khi ta phát biểu “Thà đoán lầm còn hơn bỏ sót”, là ta đang xem độ đo nào quan trọng hơn?	Đáp án đúng: Recall
Câu 89: Chọn phát biểu ĐÚNG trong các phát biểu sau?	Đáp án đúng: Khi Recall tăng thì Precision giảm
Câu 90: Dựa vào bảng sau, giá trị của xác suất $P(\text{Thu nhập} = \text{“Cao”} \text{An toàn} = \text{“Có”})$ là Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có	Đáp án đúng: 2/5
Câu 91: Dựa vào bảng sau, khi dựng cây quyết định với thuật toán ID3 để tiên đoán độ an toàn của khách hàng. Hãy lựa chọn thông tin không cần tính. Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có	Đáp án đúng: Info An toàn (D)

Câu 92: Dựa vào bảng sau, theo thuật toán Naïve Bayes, khi cần xét độ an toàn cho khách hàng {Thu nhập = "Cao", Sở hữu nhà = "Không", Tuổi = ">45", Nghề nghiệp = "CNV"}, ta KHÔNG cần tính xác suất nào? Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có	Đáp án đúng: $P(\text{An toàn} = \text{"Có"} \text{Nghề nghiệp} = \text{"CNV"})$
Câu 93: Cho các xác suất sau: $P(\text{An toàn} = \text{"Có"} \text{Nghề nghiệp} = \text{"CNV"})$ 1) $P(\text{Túổi} = \text{"40-45"} \text{An toàn} = \text{"Có"})$ 2) $P(\text{Túổi} = \text{"40-45"} \text{An toàn} = \text{"Không"})$ 3) $P(\text{An toàn} = \text{"Không"} \text{Nghề nghiệp} = \text{"CNV"})$ 4) $P(\text{An toàn} = \text{"Có"} \text{Nghề nghiệp} = \text{"CNV"})$ 5) $P(\text{Thu nhập} = \text{"Cao"} \text{An toàn} = \text{"Có"})$ 6) $P(\text{Thu nhập} = \text{"Cao"} \text{An toàn} = \text{"Không"})$ 7) $P(\text{An toàn} = \text{"Có"} \text{Thu nhập} = \text{"Cao"})$ 8) $P(\text{An toàn} = \text{"Không"} \text{Thu nhập} = \text{"Cao"})$ 9) $P(\text{An toàn} = \text{"Có"})$ 10) $P(\text{An toàn} = \text{"Không"})$ Dựa vào bảng sau, theo thuật toán Naïve Bayes, hãy chọn các xác suất được sử dụng khi cần xét độ an toàn cho khách hàng {Thu nhập = "Cao", Tuổi = ">45", Nghề nghiệp = "CNV"}. Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có	Đáp án đúng: 1, 2, 5, 6, 9, 10
Câu 94: Đường mũi tên biểu thị khoảng cách gì giữa 2 cụm 	Đáp án đúng: Single - link

Câu 95: Chọn phát biểu ĐÚNG theo thuật toán DBSCAN 	Đáp án đúng: Điểm p là đến được trực tiếp theo mật độ từ q.
Câu 96: Thuật toán nào thích hợp để phân cụm tập dữ liệu như hình bên dưới? 	Đáp án đúng: DBSCAN
Câu 97: Lựa chọn nào SAI khi nói về yếu điều của thuật toán K-means?	Đáp án đúng: Không chắc tìm được đủ k cụm
Câu 98: Tiêu chí của phân cụm là	Đáp án đúng: Cực đại hóa độ tương đồng nội cụm và cực tiểu hóa độ tương đồng liên cụm.
Câu 99: Kết quả của thao tác phân cụm KHÔNG phụ thuộc vào	Đáp án đúng: Thời gian thực hiện thuật toán phân cụm.
Câu 100: Hình thành cụm bằng cách lan truyền theo mật độ là đặc điểm của thuật toán nào?	Đáp án đúng: DBSCAN
Câu 101: Chọn phát biểu ĐÚNG khi so sánh thuật toán PageRank và thuật toán HITS	Đáp án đúng: Thứ hạng trang web trong thuật toán PageRank được tính KHÔNG phụ thuộc vào câu truy vấn, còn thứ hạng trang web trong thuật toán HITS thì phụ thuộc vào câu truy vấn.
Câu 102: Chức năng tư vấn người dùng mua hàng trên web thương mại điện tử là kết quả của khai thác	Đáp án đúng: Lịch sử sử dụng web
Câu 103: Thao tác tính hạng trang web trong máy tìm kiếm là kết quả của khai thác	Đáp án đúng: Cấu trúc web

Câu 104: Hai bước chính của máy tìm kiếm thực hiện là gì?	Đáp án đúng: Truy hồi thông tin và tính hạng trang web.
Câu 105: Ký hiệu $a(i)$ và $h(i)$ lần lượt là authority và hub của trang i. Hãy chọn công thức đúng theo thuật toán HITS để tính hạng cho trang 1 theo lược đồ bên dưới 	Đáp án đúng: $h(1) = a(5) + a(6) + a(7)$; $a(1) = h(2) + h(3) + h(4)$
Câu 106: Nếu sử dụng thuật toán PageRank để phân hạng các trang web liên kết như đồ thị bên dưới, giả sử khởi đầu 4 nút đều có điểm pagerank là $1/4$. Sau 3 lần lặp, nút nào sẽ có điểm pagerank cao nhất? 	Đáp án đúng: d
Câu 107: Vì sao một tổ chức, công ty cần phải khai thác dữ liệu?	Đáp án đúng: Vì tổ chức đó muốn rút trích tri thức từ nguồn dữ liệu có sẵn.
Câu 108: Trong quy trình khám phá tri thức, khai thác dữ liệu là bước nằm ngay sau thao tác	Đáp án đúng: Tiền xử lý dữ liệu
Câu 109: Khái niệm nào sau đây không có liên quan đến khai thác dữ liệu?	Đáp án đúng: Phân tích nghiệp vụ (business analysis)
Câu 110: Hãy chọn phát biểu ĐÚNG trong các phát biểu sau:	Đáp án đúng: Khai thác dữ liệu là một bước tiến của khoa học dữ liệu.
Câu 111: Một siêu thị muốn triển khai một số chính sách khuyến mãi cho khách hàng. Siêu thị muốn chính sách khuyến mãi của mình phù hợp nhất với nhu cầu của khách hàng. Vậy, siêu thị nên thực hiện bài toán gì trên thông tin và lịch sử mua hàng của khách hàng?	Đáp án đúng: Phân cụm
Câu 112: Trong một công ty, ai là người sử dụng kết quả của khai thác dữ liệu?	Đáp án đúng: Bộ phận ra quyết định nghiệp vụ

Câu 113: Bạn phân tích dữ liệu về dân số của một địa phương, sau đó bạn muốn dự đoán tỷ lệ sinh của địa phương đó trong năm tới. Bạn dùng bài toán	Đáp án đúng: Hồi quy
Câu 114: Việc khai thác mối quan hệ bạn bè trên mạng xã hội sử dụng bài toán	Đáp án đúng: Khai thác đồ thị
Câu 115: Bùng nổ dữ liệu không liên quan đến	Đáp án đúng: Giả thành phần mềm
Câu 116: Tại sao nói khai thác dữ liệu được xem là là một hợp lưu của nhiều lĩnh vực?	Đáp án đúng: Vì khai thác dữ liệu sử dụng kiến thức, kỹ thuật của nhiều lĩnh vực khác nhau.
Câu 117: Câu hỏi “Ta nên duyệt cho khách hàng có thông tin là X vay bao nhiêu tiền?” có thể trả lời bằng:	Đáp án đúng: Hồi quy
Câu 118: Câu hỏi “Liệu doanh thu của công ty tăng hay giảm trong 3 tháng kế tiếp?” có thể được trả lời bằng:	Đáp án đúng: Phân tích dữ liệu chuỗi thời gian
Câu 119: Để xác định đặc trưng của các đối tượng khách hàng ta sử dụng bài toán	Đáp án đúng: Phân cụm
Câu 120: Trung vị (median) của một tập dữ liệu là	Đáp án đúng: Giá trị chính giữa của tập dữ liệu đó
Câu 121: Chiều cao, cân nặng là dữ liệu kiểu	Đáp án đúng: Số liên tục
Câu 122: Hãy cho biết hình bên dưới được gọi là gì? 	Đáp án đúng: Boxplot
Câu 123: Để đánh giá mức độ đáng tin của giá trị trung bình (mean) của một tập dữ liệu, ta cần xem xét thêm giá trị	Đáp án đúng: Độ lệch chuẩn

Câu 124: Dữ liệu nào là dữ liệu nhị phân đối xứng?	Đáp án đúng: Giới tính (+/-)
Câu 125: $d(i, j) = \sqrt[h]{ x_{i1} - x_{j1} ^h + x_{i2} - x_{j2} ^h + \dots + x_{ip} - x_{jp} ^h}$ là công thức tính khoảng cách	Đáp án đúng: Minkowski
Câu 126: Hai kiểu lược đồ thường sử dụng trong data warehouse là	Đáp án đúng: Lược đồ hình sao và lược đồ hình bông tuyết.
Câu 127: Hãy lựa chọn đặc trưng đúng của OLAP khi so sánh với OLTP	Đáp án đúng: OLAP thường có truy vấn phức tạp hơn OLTP.
Câu 128: Trình tự các bước triển khai Data Mart	Đáp án đúng: Thiết kế, Xây dựng, Cư trú, Truy cập, Quản lý.
Câu 129: OLAP là tên viết tắt của kĩ thuật nào?	Đáp án đúng: Online Analytical Processing
Câu 130: Hãy chọn phát biểu ĐÚNG khi so sánh thuật toán Apriori và thuật toán FP-Growth	Đáp án đúng: FP-Growth không sinh tập ứng viên như Apriori.
Câu 131: Độ đo về tính dễ bắt gặp của luật kết hợp được gọi là	Đáp án đúng: Độ hỗ trợ (support)
Câu 132: Trong các khuyết điểm sau đây, khuyết điểm nào KHÔNG phải của thuật toán Apriori?	Đáp án đúng: Không sinh tập ứng viên
Câu 133: Theo cơ sở dữ liệu giao tác trong bảng sau, độ hỗ trợ của tập AB là: TID Itemset 100 A B C D 200 A C D 300 A B D 400 B C D E 500 A B C E	Đáp án đúng: 60%
Câu 134: Theo cơ sở dữ liệu giao tác trong bảng sau, độ hỗ trợ và độ tin cậy của luật kết hợp BC à D lần lượt là: TID Itemset 100 A B C D 200 A C D 300 A B D 400 B C D E 500 A B C E	Đáp án đúng: 40%; 66.7%
Câu 135: Hãy cho biết phát biểu nào sau đây ĐÚNG với tính chất downward closure?	Đáp án đúng: Mọi tập bao của một tập không phổ biến thì không phổ biến.
Câu 136: Ưu điểm của thuật toán FP-Growth là	Đáp án đúng: Không sinh tập ứng viên

Câu 137: Thuật toán FP-Growth cần nhiều bộ nhớ để thực thi do	Đáp án đúng: Phải nén toàn bộ cơ sở dữ liệu giao tác vào cấu trúc cây FP.
Câu 138: Hãy chọn cụm từ nào sau đây có liên quan đến thuật toán FP-Growth?	Đáp án đúng: Đệ quy
Câu 139: Để xác định khách hàng thường mua gì sau khi mua siêu xe Rolls Royce, ta làm bài toán khai thác	Đáp án đúng: Luật kết hợp hiếm
Câu 140: Cho X là một tập mục, X là tập phổ biến cực đại (max-pattern) khi và chỉ khi X là phổ biến và	Đáp án đúng: Không tồn tại một tập mục Y sao cho $Y \subset X$, mà Y là phổ biến
Câu 141: Trong các ứng dụng sau, ứng dụng nào KHÔNG PHẢI là ứng dụng của phân lớp?	Đáp án đúng: Phân tích giỏ hàng
Câu 142: Công thức sau được ứng dụng trong thuật toán nào? $P(H X) = \frac{P(X H)P(H)}{P(X)}$	Đáp án đúng: Naïve Bayes
Câu 143: Độ đo Information Gain được sử dụng để	Đáp án đúng: Lựa chọn thuộc tính để rẽ nhánh
Câu 144: Thuật toán nào KHÔNG áp dụng được cho dữ liệu trong bảng sau nếu không xử lý gì thêm? Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có	Đáp án đúng: KNN
Câu 145: Thuật toán nào sau đây cần giả định các thuộc tính của dữ liệu là độc lập về mặt thống kê?	Đáp án đúng: Naïve Bayes
Câu 146: Overfitting là hiện tượng xảy ra khi	Đáp án đúng: Sử dụng quá nhiều dữ liệu để huấn luyện mô hình.

Câu 147: Độ chính xác (accuracy) của phép tiên đoán trong bảng 2 là: Actual\Prediction Cancer Not Cancer Total Cancer 1300 1200 2500 Not Cancer 2700 4800 7500 Total 4000 6000 10000	Đáp án đúng: 61%
Câu 148: Tại sao nói phân lớp là phương pháp học có giám sát?	Đáp án đúng: Vì có thể kiểm định tính chính xác của mô hình và huấn luyện lại.
Câu 149: Dựa vào bảng sau, theo thuật toán Naïve Bayes, khi cần xét độ an toàn cho khách hàng {Thu nhập = "Thấp", Sở hữu nhà = "Không", Tuổi = "30-45", Nghề nghiệp = "CNV"}, ta KHÔNG cần tính xác suất nào? Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có	Đáp án đúng: $P(\text{An toàn} = \text{"Có"} \text{Nghề nghiệp} = \text{"CNV"})$
Câu 150: Trong bài toán dự đoán khách hàng tiềm năng để tiếp thị. Một khách hàng là không tiềm năng, nhưng mô hình dự đoán là tiềm năng. Khi đó doanh nghiệp sẽ...	Đáp án đúng: Mất chi phí
Câu 151: Dựa vào bảng sau, giá trị của xác suất $P(\text{Nghề nghiệp} = \text{"CNV"} \text{An toàn} = \text{"Không"})$ là: Thu nhập Sở hữu nhà Tuổi Nghề nghiệp An toàn 2 Không 30-45 CNV Có Cao Không <30 CNV Không Cao Có 30-45 Kinh doanh Không Rất cao Có >45 Kinh doanh Có Rất cao Có <30 Kinh doanh Không Cao Có >45 Kinh doanh Có 2 Có >45 Nông dân Có 2 Không >45 Nông dân Không Cao Không 30-45 CNV Có	Đáp án đúng: 1/4
Câu 152: K-means là phương pháp phân cụm dựa trên	Đáp án đúng: Phân hoạch (partitioning)
Câu 153: Trong thuật toán phân cụm dựa trên phân cấp top-down (thuật toán DIANA), với tập dữ liệu có N điểm, kết quả cuối cùng sẽ là:	Đáp án đúng: N cụm

Câu 154: Cho các bước sau: B1. Gán các điểm dữ liệu vào các cụm theo trọng tâm gần nhất B2. Chọn k trọng tâm B3. Tính lại trọng tâm mới B4. Tính khoảng cách từ các điểm dữ liệu đến k trọng tâm Thứ tự đúng của các bước theo thuật toán k-means lần lượt là:	Đáp án đúng: B2 – B4 – B1 – B3
Câu 155: Ứng dụng nào sau đây KHÔNG phải là ứng dụng của phân cụm?	Đáp án đúng: Phát hiện thư rác
Câu 156: Đối tượng của khai thác dữ liệu web bao gồm:	Đáp án đúng: Cấu trúc web, nội dung web và lịch sử sử dụng web.
Câu 157: Mục đích chính của khai thác cấu trúc web là tìm ra những mối quan hệ chưa biết giữa	Đáp án đúng: Các trang web
Câu 158: Chọn phát biểu ĐÚNG trong thuật toán HITS	Đáp án đúng: Authority của trang p cao khi p được nhiều trang có hub cao trỏ đến