1. 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?

1. 24 (chẵn nên (24+24)/2
2. 25
3. 20
4. 22.5
5. 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?

1. 4.71
2. 16.52
3. 6
4. 4.23
5. 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
6. Trung vị
7. Yếu vị (mode)
8. Độ lệch chuẩn
9. Cực đại
10. Hãy cho biết các chỉ số thể hiện trong biểu đồ boxplot là gì?
11. Min, Max, Q1, Q3, Median
12. Min, Max, Q1, Q3, Mean
13. Q1, Q2, Q3, Q4, Q5, Mean
14. Q1, Q2, Q3, Mean, Median
15. Theo cơ sở dữ liệu giao tác trong bảng 1, độ hỗ trợ của tập AB là
16. Độ hỗ trợ = (Số lần xuất hiện của tập hỗ trợ AB) / (Tổng số giao tác) \* 100%

|  |  |
| --- | --- |
| **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 |

1. 60%
2. 75%
3. 80%
4. 65%
5. Theo cơ sở dữ liệu giao tác trong bảng 1, 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 |

1. 75%
2. 60%
3. 65%
4. 80%
5. Theo cơ sở dữ liệu giao tác trong bảng 1, 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 |

1. 40%; 66.7%
2. 66.7%; 40%
3. 40%; 75%
4. 75%; 40%
5. Theo cơ sở dữ liệu giao tác trong bảng 1, độ hỗ trợ và độ tin cậy của luật kết hợp BC 🡪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 |

1. 40%; 66.7%
2. 66.7%; 40%
3. 40%; 75%
4. 75%; 40%
5. Chọn công thức đúng để tính độ tin cậy của luật
6. 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?
7. *C3= {ABC, ABD, ACD, BCD}*
8. *C3= {ABC, ABD, BCD}*
9. *C3= {ABD, BCD}*
10. *C3 = {BCD*}
11. Trung vị (median) của một tập dữ liệu là
12. Giá trị trung bình của tập dữ liệu đó
13. Giá trị chính giữa của tập dữ liệu đó
14. Giá trị cao nhất của tập dữ liệu đó
15. Giá trị xuất hiện nhiều nhất của trong tập dữ liệu
16. **Vì sao một tổ chức cần phải khai thác dữ liệu?**

A. Vì tổ chức đó có quá nhiều dữ liệu

B. Vì tổ chức đó thiếu lao động có tri thức

C. Vì cơ sở dữ liệu của tổ chức đó không được thiết kế tốt

D. Vì tổ chức đó muốn rút trích tri thức từ nguồn dữ liệu có sẵn

1. **Khai thác dữ liệu có thể giúp cho doanh nghiệp…**

A. Thu thập dữ liệu hiệu quả hơn

B. Cải thiện hiệu năng hoạt động của cơ sở dữ liệu

C. Phát hiện và khắc phục điểm yếu của cơ sở dữ liệu

D. Cải thiện chiến lược kinh doanh

1. **Trong quy trình khám phá tri thức, khai thác dữ liệu là bước nằm ngay sau thao tác...**

A. Thu thập dữ liệu

B. Tiền xử lý dữ liệu

C. Trực quan hóa dữ liệu

D. Tích hợp dữ liệu

1. **Tri thức khám phá được từ quy trình khai thác dữ liệu được sử dụng bởi ai?**

A. Chuyên viên phân tích số liệu

B. Chuyên viên quản trị cơ sở dữ liệu

C. Chuyên viên hoạch định chiến lược

D. Chuyên viên quản trị mạng

1. **OLAP nghĩa là?**

A. Phân tích dữ liệu trực tuyến

B. Phân tích dữ liệu ngoại tuyến

C. Phân tích dữ liệu trực quan

D. Phân tích dữ liệu trực tiếp

1. **Vấn đề nào sau đây KHÔNG phải là khó khăn(thách thức) trong khai phá dữ liệu?**

A. Có quá nhiều dữ liệu

B. Phải tích hợp dữ liệu từ nhiều nguồn

C. Dữ liệu có quá nhiều định dạng

D. Có quá nhiều thuật toán

1. **Phân tích giỏ hàng là bài toán ...**

A. Tìm luật kết hợp

B. Phân lớp

C. Phân cụm

D. Hồi quy

1. **Hãy chọn phát biểu ĐÚNG trong các phát biểu sau:**

A. Khai thác dữ liệu là một bước tiến của khoa học máy tính

B. Khai thác dữ liệu là một bước tiến của khoa học dữ liệu

C. Khai thác dữ liệu là một bước tiến của khoa học thực nghiệm

D. Khai thác dữ liệu là một bước tiến của khoa học lý thuyết

1. **Hệ thống nào sau đây sử dụng khai thác dữ liệu?**

A. Hệ thống thông tin quản lý

B. Hệ thống hỗ trợ quyết định

C. Hệ thống thông tin địa lý

D. Hệ thống thông tin phân tán

1. **Lọc thư rác là ứng dụng của bài toán ...**

A. Phân cụm

B. Phân lớp

C. Hồi quy

D. Phân tích luật kết hợp

1. **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?**

A. Phân cụm

B. Phân lớp

C. Hồi quy

D. Phân tích luật kết hợp

1. **Giai đoạn nào được thực hiện đầu tiên trong quy trình phát hiện tri thức?**

A. Khai thác dữ liệu

B. Đánh giá tri thức

C. Làm sạch dữ liệu

D. Tích hợp dữ liệu

1. **Thao tác nào sau đây không thuộc giai đoạn tiền xử lý dữ liệu?**

A. Làm sạch dữ liệu

B. Trực quan hóa dữ liệu

C. Rút gọn dữ liệu

D. Chuẩn hóa dữ liệu

1. **Tích hợp dữ liệu nhằm ...**

A. Gom những mẫu dữ liệu giống nhau vào một nhóm

B. Gom những mẫu dữ liệu cùng nằm vào một khối

C. Tập hợp dữ liệu từ nhiều nguồn thành một khối

D. Tập hợp dữ liệu cũ và dữ liệu mới thành môt khối

1. **Một trong những kỹ thuật để phát hiện ngoại lệ (outlier detection) là sử dụng bài toán**

A. Phân cụm

B. Phân lớp

C. Phân tích chuỗi

D. Khai thác luật kết hợp

1. **Việc khai thác mối quan hệ bạn bè trên mạng xã hội sử dụng bài toán**

A. Khai thác đồ thị

B. Khai thác hình ảnh

C. Khai thác văn bản

D. Khai thác đa phương tiện

1. **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**

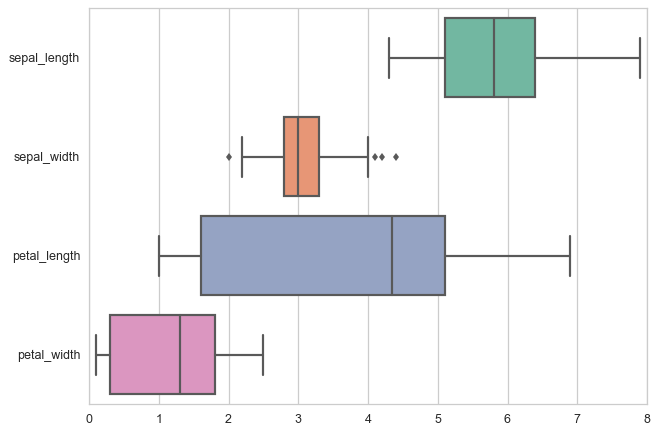
A. Học không giám sát

B. Học có giám sát

C. Phân tích tương quan

D. Phân tích luật kết hợp

1. **Hãy cho biết hình bên dưới được gọi là gì?**



A. Boxplot

B. Scatter plot

C. Histogram

D. Bar chart

1. **Đối với thuật toán Apriori, để tìm tất cả các luật kết hợp, ta cần phải làm gì trước đó?**

A. Tìm tất cả các tập phổ biến

B. Tìm tất cả các tập không phổ biến

C. Tính độ hỗ trợ của các mục đơn

D. Tính độ tin cậy của các mục đơn

1. **Độ đo về tính dễ bắt gặp của luật kết hợp được gọi là**

A. Độ tin cậy (confidence)

B. Độ hỗ trợ (support)

C. Độ nâng (lift)

D. Độ nhạy (sensitivity)

1. **Ứng dụng nào sau đây KHÔNG áp dụng bài toán khai thác tập phổ biến?**

A. Phân tích giỏ hàng

B. Phân tích lịch sử duyệt web (web logs)

C. Tiếp thị chéo (cross marketing)

D. Xác định hạn mức tín dụng

1. **Thuật toán FP-Growth cần quét cơ sở dữ liệu bao nhiêu lần?**

A. 1 lần

B. 2 lần

C. 3 lần

D. Không xác định

**34. Hãy cho biết trong các đặc điểm sau, đặc điểm nào KHÔNG phải của thuật toán Apriori?**

A. Cần xây dựng cây

B. Duyệt cơ sở dữ liệu nhiều lần

C. Sinh ra nhiều ứng viên

D. Sử dụng tính chất downward closure

**35. Độ hỗ trợ (support) của một luật kết hợp là**

A. Mức độ đáng tin của luật

B. Mức độ phổ biến của luật

C. Mức độ đúng đắn của luật

D. Mức độ hỗ trợ người dùng của luật

**36. Ưu điểm của thuật toán FP-Growth là**

A. Duyệt đệ quy cây FP

B. Nén cơ sở dữ liệu vào cây FP

C. Không sinh tập ứng viên

D. Tốn ít bộ nhớ để duy trì cây FP

**37. Câu nào sau đây KHÔNG đúng về Apriori khi so sánh Apriori và FP-Growth?**

A. Apriori dễ cài đặt hơn FP-Growth

B. Apriori quét cơ sở dữ liệu nhiều lần

C. Apriori gặp khó khăn khi duyệt cây đệ quy

D. Apriori gặp khó khăn khi đếm độ hỗ trợ của các tập mục

**38. Trong các ứng dụng sau, ứng dụng nào KHÔNG PHẢI là ứng dụng của phân lớp?**

A. Nhận dạng khuôn mặt

B. Nhận dạng chữ viết

C. Phát hiện thư rác

D. Phân tích giỏ hàng

**39. Thuật toán nào sau đây sử dụng Entropy làm độ đo thông tin?**

A. ID3

B. KNN

C. Naïve Bayes

D. SVM

**40. Khái niệm nào sau đây không có liên quan đến khai thác dữ liệu?**

1. Phát hiện tri thức (Knowledge discovery)
2. Rút trích tri thức (knowledge extraction)
3. Nghiệp vụ thông minh (business intelligence)
4. Phân tích nghiệp vụ (business analysis)

**41. Hệ thống nào sau đây sử dụng khai thác dữ liệu ?**

1. Hệ thống thông tin quản lý
2. Hệ thống hỗ trợ quyết định
3. Hệ thống thông tin địa lý
4. Hệ thống thông tin phân tán

**42. Bùng nổ dữ liệu không liên quan đến**

1. Năng lực của máy tính
2. Định luật Moore
3. Giá thành phần cứng
4. Giá thành phần mềm

**43. Hệ thống khuyến nghị (recommender systems) trong tiếp thị chéo (cross-marketing) thường sử dụng kết quả của**

1. Luật kết hợp
2. Phân cụm
3. Hồi quy
4. Phân lớp

**44. Vẽ biểu đồ cho dữ liệu ta gọi là**

1. Trực quan hóa dữ liệu
2. Tiền xử lý dữ liệu
3. Chuyển dạng dữ liệu
4. Chuẩn hóa dữ liệu