**Ôn tập thi kết thúc học phần môn Khai phá dữ liệu**

Thời gian bài thi: 75p

Đề thi bao gồm 2 câu hỏi

Câu 1:

* Chương 1, 2, 3
* 5 ý
* 6 điểm

Câu 2: Các kỹ thuật khai phá dữ liệu

* 2 ý
* Các dạng bài:
  + Khai phá luật kết hợp: Apriori, FP-Growth
  + Phân lớp dữ liệu: ID3, C4.5
  + Phân cụm dữ liệu: K-means

1. **Câu 1:**

* Đề thi mẫu:

A screenshot of a paper with text

Description automatically generated

1. **Câu 2:**
   1. **Khai phá luật kết hợp: Apriori, FP-Growth**

A notebook with writing on it

Description automatically generated

Bài tập:

A table with numbers and a list of items

Description automatically generated

* + 1. Tìm các tập phổ biến
    2. Xác định các luật kết hợp mạnh
  1. Apriori:
  2. FP-Growth:

1. **Phân lớp dữ liệu: ID3, C4.5**
   1. ID3:

A white sheet with writing on it

Description automatically generated

A notebook with math equations

Description automatically generated

A notebook with writing on it

Description automatically generated

A notebook with math equations

Description automatically generated

* 1. C4.5:

A close-up of a chart

Description automatically generated

- Thông tin dự kiến cần thiết để phân loại một bộ dữ liệu trong D được cung cấp bởi:

A mathematical equation with numbers and symbols

Description automatically generated = log2 log2 = 0.940 bits

- Tính thông tin cho từng thuộc tính :

Infoage(D) = x Info(Dj) = log2 log2) +  log2 )

+ log2 log2)= 0.694 bits.

- Tương tự: Infoincome(D) = 0.911 bits.

Infostudent(D) = 0.789 bits.

Infocredit rating(D) = 0.892 bits.

- Do đó, việc thu được thông tin từ việc phân vùng như vậy sẽ là:

Gain(age) = Info(D) − Infoage(D) = 0.940 − 0.694 = 0.246 bits

- Tương tự : Gain(income) = 0.029 bits,

Gain(student) = 0.151 bits,

Gain(credit rating) = 0.048 bits

- Giá trị “tách thông tin” được xác định tương tự với Thông tin (D) là :

SplitInfoage (D)= log2 = log2 log2 log2 =1.577 bits.

Tương tự : SplitInfoincome(D) = 1.557 bits.

SplitInfostudent(D) = 1 bits.

SplitInfocredit rating(D) = 0,985 bit

- Giá trị “tách thông tin” được xác định tương tự với Thông tin (D) là :

- Tỷ lệ đạt được được là :

GainRatio(age) = = = 0.156

Tương tự :

GainRatio(income) = = = 0.019

GainRatio(student) = = = 0.151

GainRatio(age) = = = 0.049

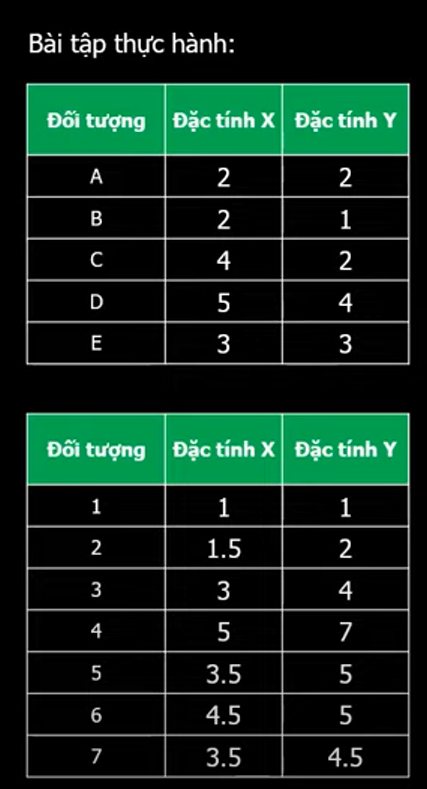
=> tỉ lệ của thuộc tính age là cao nhất => Thuộc tính age là tốt nhất để trở thành cây.

Làm tương tự các nhánh ta có được cây quyết định:

A diagram of a person's age

Description automatically generated

1. **Phân cụm dữ liệu: K-means**

****

**Video cách làm:**

https://www.youtube.com/watch?v=bIioxQE1ow4&ab\_channel=TanoT%C3%A2n