Câu 1:

1. Tính các giá trị

Trung bình:

A math equation with numbers and symbols

Description automatically generated

Trung vị:

* Sắp xếp trước khi tính
* N chẵn:

Lấy tổng 2 số giữa + vào chia đôi

* N lẻ:

Lấy luôn số chính giữa

Mode:

Số xuất hiện nhiều nhất nếu có nhiều hơn 1 cái

# b) Vẽ biểu đồ Boxplot

tính:

min

max

Vị trí các phân vị:

*Nếu ra số thập phân thì lấy trung bình giá trị của 2 vị trí trước và sau; nếu ra số nguyên thì là giá trị tại vị trí đó*

* Q1: .
* Q2: median
* Q3:

Tính IQR để tìm min max và giá trị ngoại lệ

IQR= Q3 - Q1

Khoảng giá trị hợp lệ: [ Q1 – 1,5.IQR ; Q3 + 1,5.IQR ]

# c, Chuẩn hoá

**Chuẩn hoá Min – Max**:



**Chuẩn hoá Z-score**:



SA = 

**Chuẩn hoá thập phân**:



Trong đó j là giá trị nhỏ nhất để 

Cho dãy sô 1, 5, 20, 70, 80

j=2 vì 80 < 100. Do giá trị lớn nhất là 80 nên lấy 102 là 100 giá trị gần nhất lớn hơn 80 vậy j=2.

# d, Bin

**Bin theo chiều rộng**:

Xét phần tử trong bin theo độ rộng mỗi bin, lấy (max-min) chia số bin để tìm số giá trị trong mỗi bin hoặc lấy (max-min) chia cho số giá trị trong mỗi bin để tìm số bin.



Ví dụ: 5, 10, 15, 20, 50, 60, 80. 4 bin

(80-5)/4= 18,75 xấp xỉ 19. Mỗi bin có độ rộng là 19

Bin 1: [5-24) 5,15,20

Bin 2: [24-43)

Bin 3: [43-62) 50,60

Bin 4: [62-81) 80

**Bin theo chiều sâu**:

Lấy số giá trị chia số số bin để có số giá trị mỗi bin hoặc số giá trị chia số giá trị mỗi bin để tìm bin



Ví dụ: 5, 10, 15, 20, 50, 60, 80. 4 bin

Số giá trị mỗi bin 7/4 = 1.75 **Làm tròn lên** giá trị trong mỗi bin là 2 2 2 1

Bin 1: 5,10

Bin 2: 15,20

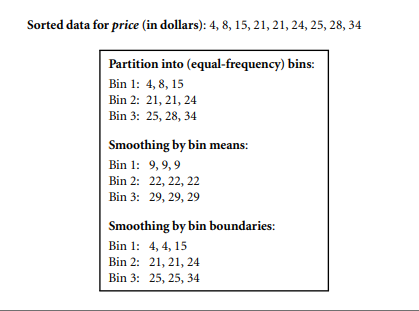
Bin 3: 50,60

Bin 4: 80

**Làm mịn:**

Làm mịn theo giá trị trung bình

Làm mịn theo theo giá trị biên



# e, Hệ số tương quan giữ X và Y





Câu 2:

**Apriori:**

A paper with writing on it

Description automatically generated

Tính |D| = 5 (Số bản ghi)

min\_sup = 60%

=> min\_sup\_count =  => Làm tròn lên

min\_sup\_count = 3

**Thuật toán Apriori:**

* Vòng lặp 1:

C1:

|  |  |
| --- | --- |
| Item | Sup\_count |
| I1 | 3 |
| I2 | 3 |
| I3 | 4 |
| I4 | 2 |
| I5 | 3 |
| I6 | 2 |
| I7 | 1 |

So sánh ≥ min\_sup\_count = 3

L1:

|  |  |
| --- | --- |
| Item | Sup\_count |
| I1 | 3 |
| I2 | 3 |
| I3 | 4 |
| I5 | 3 |

* Vòng lặp 2:

C2:

|  |  |
| --- | --- |
| Item | Sup\_count |
| {I1, I2} | 2 |
| {I1, I3} | 3 |
| {I1, I5} | 2 |
| {I2, I3} | 3 |
| {I2, I5} | 3 |
| {I3, I5} | 3 |

So sánh ≥ min\_sup\_count = 3

L2:

|  |  |
| --- | --- |
| Item | Sup\_count |
| {I1, I3} | 3 |
| {I2, I3} | 3 |
| {I2, I5} | 3 |
| {I3, I5} | 3 |

* Vòng lặp 3:

C3: Ghép các cặp mà cùng phần tử đầu

|  |  |
| --- | --- |
| Item | Sup\_count |
| {I2, I3, I5} | 3 |

So sánh ≥ min\_sup\_count = 3

L3:

|  |  |
| --- | --- |
| Item | Sup\_count |
| {I2, I3, I5} | 3 |

Vậy các tập phổ biến là L = L1 v L2 v L3

**FP-Growth:**

|  |  |
| --- | --- |
| ID | Items |
| T1 | F, M, B, E |
| T2 | O, F, E |
| T3 | O, A, H, S |
| T4 | D, B, F, E |
| T5 | D, A, H, S |
| T6 | O, M, E |
| T7 | O, A, D, H, S |

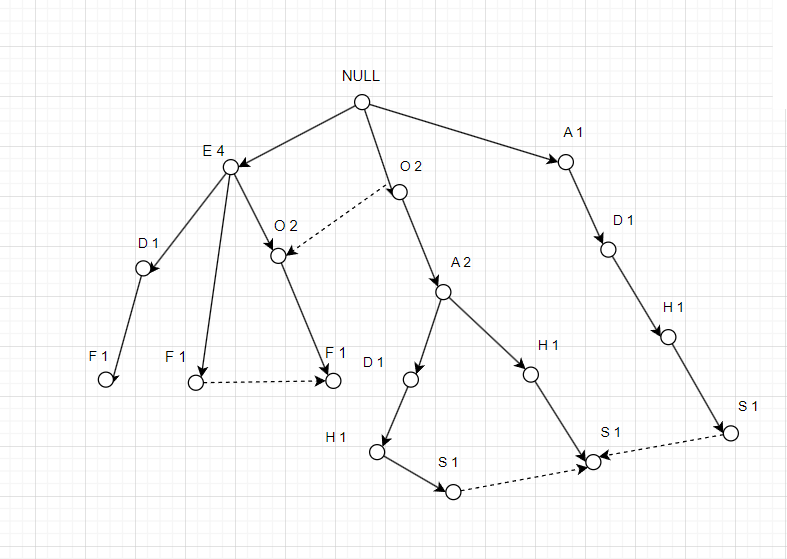
**Thuật toán FP-Growth:**

|  |  |
| --- | --- |
| Item | Sup\_count |
| A | 3 |
| B | 2 |
| D | 3 |
| E | 4 |
| F | 3 |
| H | 3 |
| M | 2 |
| O | 4 |
| S | 3 |

|  |  |
| --- | --- |
| Item | Sup\_count |
| E | 4 |
| O | 4 |
| A | 3 |
| D | 3 |
| F | 3 |
| H | 3 |
| S | 3 |

|  |  |
| --- | --- |
| ID | Items |
| T1 | E, F |
| T2 | E, O, F |
| T3 | O, A, H, S |
| T4 | E, D, F |
| T5 | A, D, H, S |
| T6 | E, O |
| T7 | O, A, D, H, S |

VẼ CÂY:

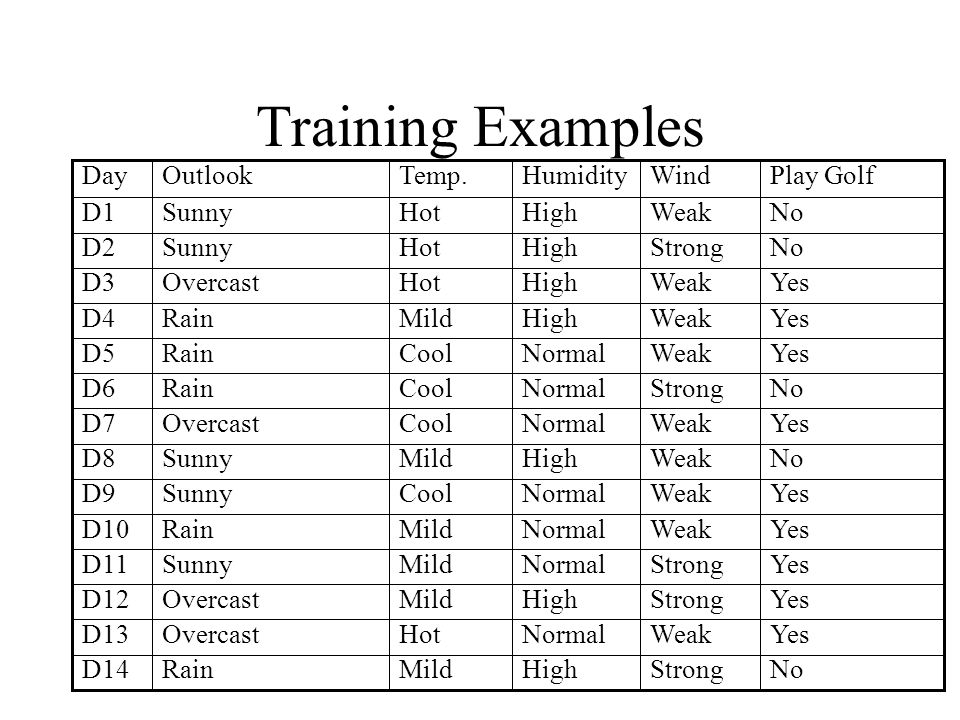


|  |  |
| --- | --- |
| Item | Conditional pattern base |
| S | {{O, A, D, H: 1}, {O, A, H: 1}}, {{A, D, H: 1}} |
| H | {{O, A, D: 1}, {O, A: 1}}, {{A, D:1}} |
| F | {{E, D: 1}, {E: 1}, {E, O: 1}} |
| D | {{E: 1}}, {{A: 1}} |
| A | {{O: 2}} |
| O | {{E: 2}} |
| E | NULL |

|  |  |
| --- | --- |
| Item | Conditional Frequent PatternTree |
| S | {{A, H: 3}} |
| H | {{A: 3}} |
| F | {{E: 3}} |

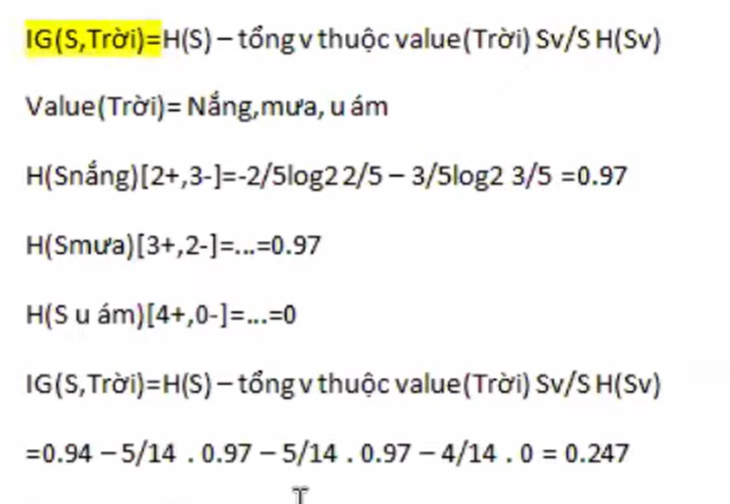
|  |  |
| --- | --- |
| Item | Frequent Pattern Generated |
| S | {{S, H: 3}, {A, S: 3}, {A, H, S: 3}} |
| H | {{A, H: 3}} |
| F | {{F, E: 3}} |

C4.5:



* Xét tập dữ liệu có: 5 N0, 9 YES:
* Info(D) = -5/14 log2(5/14) – 9/14log2(9/14) = 0,94
* Xét thuộc tính Outlook:
* InfoOutlook(D) = 5/14(-3/5log2(3/5)-2/5log2(2/5)) \*2 + 4/14 \*0

= 0.6935



* Tương tự ta có:
* Xét thuộc tính Temp:
* InfoTemp = 4/14(-2/4log2(2/4)-2/4log2(2/4)) + 6/14(-4/6log2(4/6)-2/6log2(2/6)

+ 4/14(-3/4log2(3/4)-1/4log2(1/4))

= 0,911

* Xét thuộc tính Humidity:
* InfoHumidity = 7/14(-4/7log2(4/7)-3/7log2(3/7)) + 7/14(-6/7log2(6/7)-1/7log2(1/7))

= 0,788

* Xét thuộc tính Wind:
* InfoWind = 9/14(-2/9log2(2/9)-7/9log2(7/9)) + 5/14(-2/5log2(2/5)-3/5log2(3/5))

= 0,838

* Tính Gain:

GainOutlook = InfoD – InfoOutlook = 0,940 – 0,694 = 0,246

GainTemp = InfoD – InfoTemp = 0,9402 – 0,911 = 0,029

GainHumidity = InfoD – InfoHumidity = 0,9402 – 0,788 = 0,152

GainWind = InfoD – InfoWind = 0,9402 – 0,838 = 0,102

* Tính SplitInfo:

SplitInfoOutlook = - 5/14log2(5/14) – 5/14log2(5/14) – 4/14log2(4/14) = 1,577

SplitInfoTemp = - 4/14log2(4/14) – 4/14log2(4/14) – 6/14log2(6/14) = 1,557

SplitInfoHumidity = - 7/14log2(7/14) – 7/14log2(7/14) = 1

SplitInfoWind = - 9/14log2(9/14) – 5/14log2(5/14) = 0,94

* Tính GainRatio:

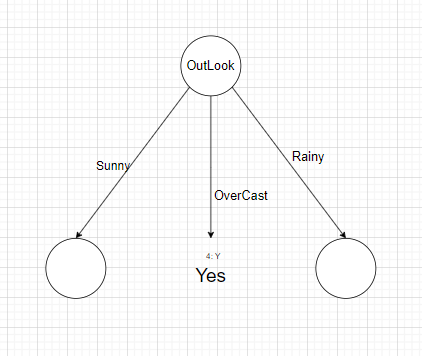
GainRatioOutlook = GainOutlook / SplitInfoOutlook = 0,246 / 1,577 = 0,156

GainRatioTemp = GainTemp / SplitInfoTemp = 0,029 / 1,557 = 0,019

GainRatioHumidity = GainHumidity / SplitInfoHumidity = 0,152 / 1 = 0,152

GainRatioWind = GainWind / SplitInfoWind = 0,102 / 0,94 = 0,109

* So sánh ta thấy GainRatioOutlook = 0,156 là lớn nhất => lấy Outlook là thuộc tính chia.



* D1 = Sunny:
* Tính Info(D1):
* Info(D1) = -2/5 log2(2/5) – 3/5log2(3/5) = 0,971
* Xét thuộc tính Temp:
* InfoTemp(D1) = 2/5 \*0(Hot) + 1/5 \*0(Cool) + 2/5 \* 1(Mild) = 0,4
* Tương tự với Humidity, Wind:

InfoHumidity(D1) = 3/5 \*0 + 2/5 \*0 = 0

InfoWind(D1) = 3/5(-2/3log2(2/3)-1/3log2(1/3)) + 2/5 \*1 = 0,951

* Tính Gain:

GainTemp = Info(D1) – InfoTemp(D1) = 0,971 – 0,4 = 0,571

GainHumidity = Info(D1) – InfoHumidity(D1) = 0,971 – 0 = 0,971

GainWind = Info(D1) – InfoWind(D1) = 0,971 – 0,951 = 0,02

* Tính SplitInfo:

SplitInfoTemp = -2/5log2(2/5) – 1/5log2(1/5) – 2/5log2(2/5) = 1,522

SplitInfoHumidity = -3/5log2(3/5) – 2/5log2(2/5) = 0,971

SplitInfoWind = -3/5log2(3/5) – 2/5log2(2/5) = 0,971

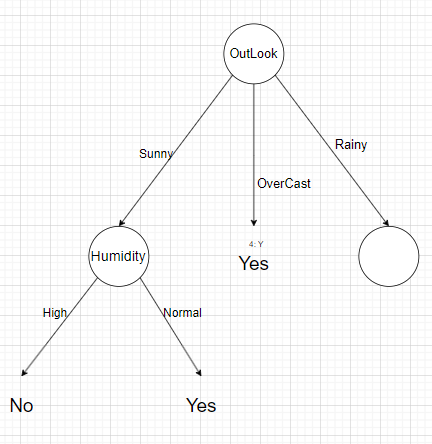
* Tính GainRatio:

GainRatioTemp = GainTemp / SplitInfoTemp = 0,571 / 1,522 = 0,375

GainRatioHumidity = GainHumidity / SplitInfoHumidity = 0,971 / 0,971 = 1

GainRatioWind = GainWind / SplitInfoWind = 0,02 / 0,971 = 0,093

* Dễ thấy GainRatioHumidity = 1 là lớn nhất => lấy Humidity là thuộc tính chia



* D2 = Rain:
* Tính Info(D2):
* Info(D2) = -2/5 log2(2/5) – 3/5log2(3/5) = 0,971
* Xét thuộc tính Temp:
* InfoTemp(D2) = 3/5(-2/3log2(2/3)-1/3log2(1/3)) + 2/5 \* 1 = 0,951
* Tương tự với Humidity, Wind:

InfoHumidity(D2) = 2/5 \*1 + 3/5(-2/3log2(2/3)-1/3log2(1/3)) = 0,951

InfoWind(D2) = 3/5 \*0 + 2/5 \*0 = 0

* Tính Gain:

GainTemp = Info(D2) – InfoTemp(D2) = 0,971 – 0,951 = 0,02

GainHumidity = Info(D2) – InfoHumidity(D2) = 0,971 – 0,951 = 0,02

GainWind = Info(D2) – InfoWind(D2) = 0,971 – 0 = 0,971

* Tính SplitInfo:

SplitInfoTemp = -2/5log2(2/5) – 3/5log2(3/5) = 0,971

SplitInfoHumidity = -3/5log2(3/5) – 2/5log2(2/5) = 0,971

SplitInfoWind = -3/5log2(3/5) – 2/5log2(2/5) = 0,971

* Tính GainRatio:

GainRatioTemp = GainTemp / SplitInfoTemp = 0,02 / 0,971 = 0,02

GainRatioHumidity = GainHumidity / SplitInfoHumidity = 0,02 / 0,971 = 0,02

GainRatioWind = GainWind / SplitInfoWind = 0,971 / 0,971 = 1

Dễ thấy GainRatioWind = 1 là lớn nhất => lấy Wind là thuộc tính chia

