**Mẫu 1:**

**Câu 1 (2 điểm)**

Cho dãy số gồm 12 số sau:

**4, 7, 7, 10, 0, 4, 7, 2, 3, 6, 5, 4**

Anh chị hãy thêm vào dãy số trên 1 số bất kỳ tùy ý, như vậy dãy số cuối cùng để thực hiện các yêu cầu dưới đây có tất cả là **13 số.**

Tính toán và ghi ra các kết quả sau:

* Mean, Min, Max, Median và Mod của dãy số này.
* Ghi ra kết quả rời rạc hóa dãy số trên nếu thực hiện rời rạc hóa bằng thuật toán Equal-Width với K=3.
* Ghi ra kết quả rời rạc hóa dãy số trên nếu thực hiện rời rạc hóa bằng thuật toán Equal-Frequency với K=3.

Yệu cầu trả lời:  Ghi vắn tắt nhưng cần rõ ràng cách thức thực hiện và kết quả của các thuật toán rời rạc hóa được thực hiện.

**Câu 1**. (***2.0 điểm***)

**Phụ thuộc vào số mà SV thêm vào (ý đồ tránh SV sao chép bài làm)**

* Mean, Min, Max, Median và Mod của dãy số (1 điểm)
* Ghi ra kết quả rời rạc hóa dãy số trên nếu thực hiện rời rạc hóa bằng thuật toán Equal-Width với K=3. (0.5 điểm)
* Ghi ra kết quả rời rạc hóa dãy số trên nếu thực hiện rời rạc hóa bằng thuật toán Equal-Frequency với K=3. (0.5 điểm)

**Bài làm:**

|  |  |
| --- | --- |
| **Mean:** | (4 + 7 + 7 + 10 + 0 + 4 + 7 + 2 + 3 + 6 + 5 + 4 + 8) / 13 = 5.3077 |
| **Min:** | 0 |
| **Max:** | 10 |
| **Median:** | **Sắp xếp dãy số:** 0, 2, 3, 4, 4, 4, 5, 6, 7, 7, 7, 8, 10  **Median:** 5 |
| **Mod:** | **Mode (Mod): Tìm số xuất hiện nhiều nhất trong dãy số.** Mode = 4 và 7 ***(vì cả hai số xuất hiện cùng số lần, không có mode đơn lẻ)*** |
| **Equal-Width K=3.** | Tính khoảng giá trị của dãy số: Max - Min = 10 - 0 = 10  Tính độ rộng của mỗi khoảng: Độ rộng = Khoảng giá trị / K = 10 / 3 ≈ 3.3333 (làm tròn đến 4 chữ số thập phân)  Xác định các khoảng giá trị:  Khoảng 1: 0 ≤ x < 3.3333  Khoảng 2: 3.3333 ≤ x < 6.6666  Khoảng 3: 6.6666 ≤ x ≤ 10  Rời rạc hóa dãy số: 0, 3, 3, 6, 0, 0, 3, 0, 0, 6, 3, 0, 6 |
| **Equal-Frequency K=3.** | Tính số lượng số trong mỗi khoảng: Số lượng số / K = 13 / 3 ≈ 4.3333 (làm tròn đến 4 chữ số thập phân)  Chia dãy số thành các nhóm có số lượng số gần nhau:  Nhóm 1: 0, 2, 3, 4  Nhóm 2: 4, 4, 5, 6  Nhóm 3: 7, 7, 7, 8, 10  Rời rạc hóa dãy số: 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3 |

**Câu 2 (2 điểm)**

Bảng dưới đây là một phần kết quả cuối năm của học sinh trường An Bình. Văn hóa được xếp loại theo 4 mức: Giỏi/Khá/Trung bình/Kém, Đạo đức (2 mức): Tốt/Kém và Thể chất (2 mức): Đạt/Không đạt

|  |  |  |  |
| --- | --- | --- | --- |
| MS | Văn Hóa={Giỏi/Khá/TB/Kém} | Đạo đức={Tốt/Kém} | Thể chất ={Đạt,Không đạt} |
| 1 | Giỏi | Tốt | Đạt |
| 2 | Khá | Tốt | Không đạt |
| 3 | Trung bình | Kém | Không đạt |
| 4 | Kém | Tốt | Đạt |
| 5 | Khá | Kém | Đạt |

Ta dự định sử dụng Luật Kết Hợp để phát hiện mối liên hệ kết quả đánh giá giữa các lĩnh vực văn hóa, đạo đức và thể chất của học sinh trường An Bình.

Ví dụ:

Văn Hóa Kém => Đạo Đức Kém, Thể Chất Không Đạt.

Văn Hóa Khá, Thể Chất Đạt => Đạo Đức Tốt.1

Anh chị hãy:

* Mô tả vắn tắt phương án chuyển đổi dữ liệu để có thể sử dụng cho thuật toán Luật Kết Hợp..
* Viết lại bảng dữ liệu trên theo phương án anh chị đã đề xuất để có thể áp dụng vào thuật toán Luật Kết Hợp.

**Câu 2**. (***2.0 điểm***)

Xây dựng bộ dữ liệu chứa các giá trị nhị phân (1/0  T/F  Y/N): VHGiỏi, VHKhá, VHTrung bình, VHKém, ĐĐTốt, ĐĐKém,  TCĐạt và TCKhông đạt (1 điểm)

|  |  |  |  |
| --- | --- | --- | --- |
| MS | Văn Hóa | Đạo đức | Thể chất ={Đạt,Không đạt} |
| 1 | Giỏi | Tốt | Đạt |
| 2 | Khá | Tốt | Không đạt |
| 3 | Trung bình | Kém | Không đạt |
| 4 | Kém | Tốt | Đạt |
| 5 | Khá | Kém | Đạt |

Kết quả chuyển đổi (1 điểm)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MS | VHGioi | VHKha | VHTrungBinh | VHKem | ĐĐTot | ĐĐKem | TCDat | TCKgDat |
| 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 3 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 4 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 5 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |

**Câu 3 (2 điểm)**

Khi giải quyết bài toán phân lớp, ta thường đánh giá mô hình dựa trên kết quả của ma trận sai lầm (confusion matrix) để tính toán các đại lượng cần thiết.

Bảng dưới đây là số liệu trong ma trận sai lầm của một mô hình :

|  |  |  |
| --- | --- | --- |
| A | B | <- Được phân lớp là |
| 40 | 10 | A |
| 15 | 35 | B |

Anh chị hãy tính toán và ghi ra các kết quả của những đại lượng theo anh chị là cần thiết để có thể đánh giá thuật toán này (làm tròn tới 2 số thập phân). Lưu ý: Cần ghi ra các số liệu cụ thể trong từng công thức tính toán và kết quả tương ứng.

Vì sao không chỉ tính toán kết quả cho cả bộ dữ liệu (bao gồm tất cả các lớp – class) mà còn phải  tính toán các kết quả cho từng lớp (class)?  Anh chị thử nêu một ví dụ minh họa và giải thích lý do.

**Câu 3**.

|  |  |  |
| --- | --- | --- |
| A | B | <- Được phân lớp là |
| 40 | 10 | A |
| 15 | 35 | B |

1 điểm: Nếu tính toán các giá trị sau

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | TP rate=  TP/(TP+FN) | FP rate=  FP/(FP+TN) | Precision=  TP/(TP+FP) | Recall=TP rate |
| Class A | 40/50=0.8 | 15/(15+35)=0.3 | 40/(40+15)=0.73 | 0.8 |
| Class B | 35/50=0.7 | 10/(10+40)=0.2 | 35/(35+10)=0.78 | 0.7 |
| Toàn bộ Data | (40+35)/100=0.75 | (10+15)/(40+35)=0.33 | (40+35)/(100)=0.75 | 0.75 |

Nếu minh họa và giải thích như sau.

Bảng 1

|  |  |  |
| --- | --- | --- |
| A | B | <- Được phân lớp là |
| 25 | 25 | A |
| 25 | 25 | B |

Bảng 2

|  |  |  |
| --- | --- | --- |
| A | B | <- Được phân lớp là |
| 0 | 50 | A |
| 0 | 50 | B |

Nếu chỉ tính trên toàn bộ dữ liệu thì TP ở cả 2 bảng đều như nhau

Bảng 1  (25+25)/100 = 50%  =  Bảng 2 (0+50)/100

Nhưng khi tính toán TP cho từng lớp thì mới thấy rõ hiệu quả của từng mô hình để sử dụng khi phân lớp.

|  |  |  |
| --- | --- | --- |
| Mô hình | TP của lớp A | TP của lớp B |
| Bảng 1 | 50% | 50% |
| Bảng 2 | 0 | 100% |

**Câu 4 (2 điểm)**

Bảng dữ liệu dưới đây lưu lại thông tin của các khách hàng vay tín dụng của ngân hàng Lạc Việt. Các thông tin lưu giữ gồm: tuổi, tình trạng gia đình, thu nhập 1 tháng (triệu đồng), số tiền vay (triệu đồng), kết quả đánh giá (Yes: Cho vay, No: Không)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Tuổi | Gia đình | Thu nhập | Số tiền vay | Kết quả |
| 1 | 25 | Có | 20 | 120 | Yes |
| 2 | 22 | Không | 15 | 200 | No |
| 3 | 24 | Có | 18 | 180 | Yes |
| 4 | 30 | Không | 17 | 130 | Yes |
| 5 | 31 | Có | 22 | 250 | No |
| 6 | 35 | Có | 22 | 150 | Yes |
| 7 | 22 | Không | 23 | 210 | No |
| 8 | 28 | Có | 19 | 280 | No |

Anh chị hãy dùng thuật toán KNN để cho biết xem có quyết định cho khách hàng có thông tin dưới đây vay tiền hay không? Anh chị toàn quyền chọn công thức tính khoảng cách và giá trị K.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tuổi | Gia đình | Thu nhập | Số tiền vay | Kết quả |
| 23 | Có | 32 | 250 | Yes/No? |

Yêu cầu: Cần trình bày vắn tắt nhưng rõ ràng các kết quả mà anh chị đã tính toán để qua đó có kết quả phân loại cuối cùng.

**GỢI Ý :**

Cần thực hiện chuẩn hóa các vùng dữ liệu trước khi tính toán khoảng cách (1 điểm).

Sử dụng giá trị K (tùy ý sinh viên) và xác định chính xác, giải thích đúng để trả lời (1 điểm).

**Bài làm:**

**Chuẩn hóa dữ liệu:**

* **Tuổi**: Không cần chuẩn hóa vì giá trị đã nằm trong khoảng nhỏ.
* **Gia đình:** Chuyển thành số (Có: 1, Không: 0).
* **Thu nhập:** Áp dụng Min-Max Scaling. ( x -min ) / (max -min)
* **Số tiền vay**: Áp dụng Min-Max Scaling.

**Tính khoảng cách:** Sử dụng khoảng cách Euclidean. **sqrt((t1 - tx)2 + (g1 - gx)2 + (tt1 – ttx)2 + (s1 – sx)2**

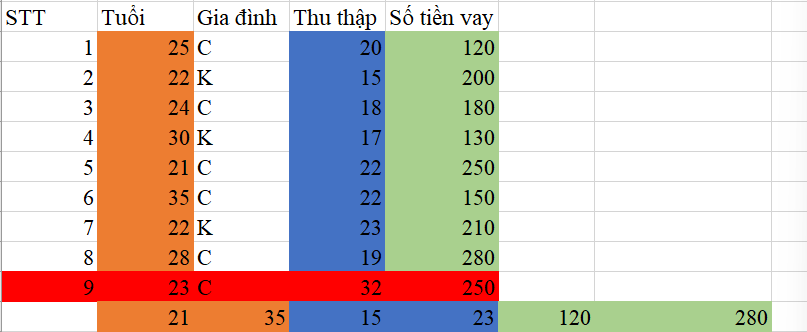
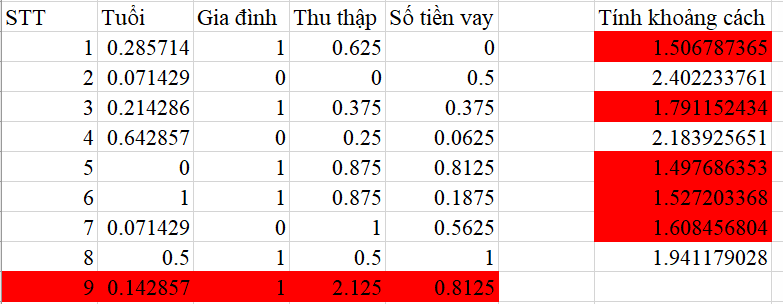
**Chọn giá trị K:** Tùy ý chọn giá trị K, ví dụ K = 5.

**Tìm các láng giềng gần nhất:**

* Tính khoảng cách từ khách hàng mới đến tất cả các khách hàng trong tập dữ liệu huấn luyện.
* Chọn 5 điểm gần nhất dựa trên khoảng cách nhỏ nhất.

**Đưa ra dự đoán:**

* Xem xét nhãn (Yes/No) của 5 láng giềng gần nhất.
* Đếm số lượng nhãn Yes và No.
* Nếu số lượng Yes lớn hơn số lượng No, khách hàng mới được dự đoán vay tiền (Yes), ngược lại, không vay tiền (No).

Vì số lượng Yes lớn hơn số lượng No, khách hàng mới được dự đoán là vay tiền (Yes).

**Câu 5 (2 điểm)**

Bài toán phân lớp đa nhãn (Multi Label Classify): là một bài toán trong Khai Phá Dữ Liệu. Trong bài toán này, dữ liệu có thể cùng lúc được gán nhiều nhãn. Ví dụ khi phân loại các bài báo, một bài báo có thể cùng lúc liên quan nhiều chủ đề:  bài báo A vừa được gán nhãn liên quan đến chủ đề Thể thao vừa gán nhãn chủ đề Thời sự, bài báo B thì chỉ được gán 1 nhãn duy nhất là chủ đề Công nghệ, bài báo C  cùng lúc được gán các chủ đề Văn hóa, Thời sự và Công Nghệ,…

Với các hiểu biết về bài toán phân lớp cổ điển (Binary và Multi Class Classify tức là bài toán phân lớp nhị phân và phân lớp với dữ liệu đa lớp) anh chị thử đề xuất một vài cách để có thể áp dụng giải quyết cho bài tóan nêu trên (phân lớp đa nhãn). Lưu ý: Trình bày vắn tắt nhưng rõ ràng, có thể minh họa bằng sơ đồ.

Hướng dẫn

* Xây dựng một loạt các mô hình phân lớp nhị phân (Binary Classify), mỗi mô hình cho phép nhận dạng 1 nhãn. Khi nhận dạng: Sẽ lần lượt thực hiện tuần tự các mô hình phân lớp nhị phân này và tổng hợp kết quả. (1 điểm)
* Chuyển bài toán phân lớp đa nhãn thành bài toán phân lớp đa lớp: mỗi lớp mới sẽ tương ứng với các tổ hợp nhãn xuất hiện trong dữ liệu. (1 điểm)

------- HẾT--------

**Mẫu 2:**

Cho bộ dữ liệu sau (đã có sẵn 9 phần tử) lưu thông tin các khách hàng đã vay tiền ngân hàng, cột kết quả (là cột để phân loại khách thành 2 nhóm) cho biết khách vay trả nợ đúng hẹn (Tốt) hay không (Xấu). Cột tuổi của khách vay. Cột gia đình cho biết khách vay có gia đình (Có) hay chưa (Không) hay ly hôn. Cột thu thập cho biết thu thập của khách hàng trong 1 tháng, cột Số tiền vay: số tiền khách muốn vay ngân hàng và 2 cột này tính theo đơn vị tính triệu đồng.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | Tuổi | Gia đình | Thu thập | Số tiền vay | Kết quả |
| 1 | 25 | C | 20 | 120 | T |
| 2 | 22 | K | 15 | 200 | X |
| 3 | 24 | L | 18 | 150 | T |
| 4 | 30 | K | 17 | 120 | T |
| 5 | 31 | C | 22 | 250 | X |
| 6 | 35 | C | 22 | 150 | T |
| 7 | 22 | K | 23 | 210 | X |
| 8 | 28 | L | 19 | 280 | X |
| 9 | 40 | C | 22 | 100 | T |
| 10 | ??? | ??? | ??? | ??? | X |

Sinh viên sẽ điền thêm thông tin cho phần tử thứ 10 (tại các ô ký hiệu ???) với thông tin tùy ý nhưng Tuổi phải trong phạm vi số nguyên từ 26 – 30, gia đình: Có hoặc Không, thu nhập là số nguyên từ 15 – 20 và tiền vay là số nguyên từ 200 – 250. Như vậy bảng dữ liệu cuối cùng sẽ có 10 phần tử. Sinh viên sẽ sử dụng dữ liệu trên (có 10 phần tử dữ liệu) để thực hiện các yêu cầu ở câu 1, 2 và 3

**Câu 1** Tính toán và ghi ra các kết quả sau:

* Mean, Min, Max, Median và Mod của Tuổi, Thu nhập và Số tiền vay.

Yêu cầu trả lời:  Ghi vắn tắt nhưng cần rõ ràng cách thức thực hiện và kết quả của các thuật toán rời rạc hóa được thực hiện.

**Tuổi (Age)**

Mean: 28.3

Min: 22

Max: 40

Median: 28

Mode: Không có giá trị mode duy nhất, vì mỗi giá trị tuổi xuất hiện một lần.

**Thu nhập (Income)**

Mean: 19.8

Min: 15

Max: 23

Median: 19

Mode: Không có giá trị mode duy nhất, vì mỗi giá trị thu thập xuất hiện một lần.

**Số tiền vay (Loan Amount)**

Mean: 173.3

Min: 100

Max: 280

Median: 150

Mode: Không có giá trị mode duy nhất, vì mỗi giá trị số tiền vay xuất hiện một lần.

* Ghi ra cách tính toàn và kết quả rời rạc hóa cột Thu nhập nếu thực hiện rời rạc hóa bằng thuật toán Equal-Width với K=3.

**Tìm giá trị nhỏ nhất và lớn nhất trong cột "Thu nhập":**

Giá trị nhỏ nhất: 15

Giá trị lớn nhất: 23

Trong trường hợp này, số khoảng rời rạc là 3, do đó khoảng giá trị cho mỗi khoảng rời rạc là (23 - 15) / 3 = 2.

**Tính toán khoảng giá trị của mỗi bin:**

Khoảng 1: [15, 17) (từ 15 đến nhỏ hơn 17)

Khoảng 2: [17, 19) (từ 17 đến nhỏ hơn 19)

Khoảng 3: [19, 23] (từ 19 đến 23)

|  |  |
| --- | --- |
| Thu thập | Thu nhập rời rạc |
| 20 | Khoảng 2 |
| 15 | Khoảng 1 |
| 18 | Khoảng 2 |
| 17 | Khoảng 1 |
| 22 | Khoảng 3 |
| 22 | Khoảng 3 |
| 23 | Khoảng 3 |
| 19 | Khoảng 2 |
| 22 | Khoảng 3 |
| 21 | Khoảng 3 |

**Câu 2:**

Với bảng dữ liệu trên (có 10 phần tử và chưa rời rạc hóa), hãy tính toán khoảng cách của các phần tử 1 và 2 (D12 ), 1 và 3 (D13 ), 2 và 3 (D23 ). Khoảng cách tính theo công thức Manhattan và chỉ dựa trên 4 cột Tuổi, Gia đình, Thu nhập và Số tiền vay.

*Lưu ý: Ghi vắn tắt, rõ ràng cách thực hiện và kết quả của các tính toán được thực hiện.*

**Phần tử 1:**

Tuổi (Age): 25

Gia đình (Family): C

Thu nhập (Income): 20

Số tiền vay (Loan Amount): 120

**Phần tử 2:**

Tuổi (Age): 22

Gia đình (Family): K

Thu nhập (Income): 15

Số tiền vay (Loan Amount): 200

**Phần tử 3:**

Tuổi (Age): 24

Gia đình (Family): L

Thu nhập (Income): 18

Số tiền vay (Loan Amount): 150

Áp dụng công thức Manhattan, ta tính khoảng cách giữa các phần tử như sau:

D12 = |25 - 22| + |C - K| + |20 - 15| + |120 - 200| = 3 + 1 + 5 + 80 = 89

D13 = |25 - 24| + |C - L| + |20 - 18| + |120 - 150| = 1 + 1 + 2 + 30 = 34

D23 = |22 - 24| + |K - L| + |15 - 18| + |200 - 150| = 2 + 1 + 3 + 50 = 56

Vậy, khoảng cách giữa các phần tử là:

Khoảng cách giữa phần tử 1 và phần tử 2 (D12): 89

Khoảng cách giữa phần tử 1 và phần tử 3 (D13): 34

Khoảng cách giữa phần tử 2 và phần tử 3 (D23): 56

**Câu 3:**

Với bảng dữ liệu (gồm 10 phần tử và chưa rời rạc hóa). Hãy rời rạc hóa các cột Tuổi, Thu nhập và Số tiền vay theo phương pháp Equal Width với K=2. Sau khi rời rạc hóa, xây dựng bảng xác suất Naïve Bayes để sau đó dự đoán xem khách hàng dưới đây thuộc nhóm tuổi nhóm Tốt và Xấu?

|  |  |  |  |
| --- | --- | --- | --- |
| Tuổi | Thu nhập | Số tiền vay | Kết quả |
| 25 | 16 | 20 | ??? |

*Lưu ý: Ghi vắn tắt nhưng cần rõ ràng cách xây dựng bằng xác suất và kết quả của các tính toán được thực hiện*

Tìm giá trị nhỏ nhất và lớn nhất của từng cột:

Cột "Tuổi": Giá trị nhỏ nhất là 22 và giá trị lớn nhất là 40.

Cột "Thu nhập": Giá trị nhỏ nhất là 15 và giá trị lớn nhất là 23.

Cột "Số tiền vay": Giá trị nhỏ nhất là 100 và giá trị lớn nhất là 280.

Tính toán khoảng giá trị của mỗi bin:

|  |  |  |
| --- | --- | --- |
| Cột "Tuổi":  Khoảng 1: (22, 31)  Khoảng 2: (31, 40) | Cột "Thu nhập":  Khoảng 1: (15, 19)  Khoảng 2: (19, 23) | Cột "Số tiền vay":  Khoảng 1: (100, 190)  Khoảng 2: (190, 280) |

Ánh xạ các giá trị ban đầu vào các bin tương ứng. Dựa trên khoảng đã tính toán, chúng ta rời rạc hóa các giá trị ban đầu.

**Với phần tử có giá trị:**

Tuổi: 25

Thu nhập: 16

Số tiền vay: 20

**Ánh xạ các giá trị vào các bin tương ứng:**

Tuổi: Khoảng 1 (22, 31)

Thu nhập: Khoảng 1 (15, 19)

Số tiền vay: Khoảng 1 (100, 190)

|  |  |  |  |
| --- | --- | --- | --- |
| Tuổi | Thu nhập | Số tiền vay | Kết quả |
| Khoảng 1 | Khoảng 1 | Khoảng 1 |  |

**Câu 4:**

Bảng dưới đây là số liệu trong ma trận sai lầm sau khi sử dụng bộ dữ liệu kiểm tra của một mô hình phân lớp:

|  |  |  |  |
| --- | --- | --- | --- |
| A (dự đoán) | B (dự đoán) | C (dự đoán) | <= Được phân lớp là |
| 40 | 5 | 5 | A (Thực tế) |
| ??? | 30 | ??? | B (Thực tế) |
| 20 | 20 | 10 | C (Thực tế) |

Anh chị hay tự điền vào 2 ô (có ký hiệu ??) các số nguyên trong phạm vi 5 – 10

Hãy tính giá trị Sensitivity và Precision của lớp A

Hãy tính TPR và FPR của cả bộ dữ liệu kiểm tra

Khi sử dụng mô hình này dự báo 1 phần tử mới thì kết quả báo phần tử mới là lớp B. Cho biết có thể tin tưởng kết quả dự đoán này bao nhiêu %?

*Lưu ý: Ghi vắn tắt nhưng cần rõ ràng các tính toán thực hiện*

**Câu 5:**

Thế nào là bài toán phân lớp đa nhãn (Multi – Label Classify)

Cho 1 ví dụ thực tế minh họa về bài toán phân lớp đa nhãn?

Với các hiểu biết về một số thuật toán giải toán phân lớp nhị nhân (Binary Classify) anh chị thử đề xuất một thuật toán có thể áp dụng giải quyết cho bài toán phân lớp đa nhãn?

*Lưu ý: Trình bày vắn tắt nhưng rõ ràng, nên minh họa bằng sơ đồ*