DÀN Ý ĐỀ TÀI CÁ NHÂN

MÔN HỌC KHOA HỌC DỮ LIỆU

1. **Phần 1: Giới thiệu về CSDL sử dụng cho đề tài**
   - Tổng quan về CSDL: dữ liệu nói về vấn đề gì? Thu thập từ nguồn nào? Table gồm bao nhiêu records? ...
   - Giới thiệu từng thuộc tính (field) có trong dữ liệu (nếu dữ liệu gồm nhiều table cũng cần liệt kê chi tiết theo từng table), gồm:
     - Tên field?
     - Ý nghĩa của field trong dữ liệu?
     - Có bao nhiêu giá trị null (missing data)?
     - Có bao nhiêu giá trị unique?
     - Kiểu dữ liệu của field là gì?
       - Nếu là kiểu **nhị phân** hay **rời rạc**: gồm các giá trị gì? Tỷ lệ phần trăm của từng giá trị? Từ đó suy ra giá trị của mode.

         Ví dụ: đối với thuộc tính nhị phân (có hút thuốc lá)

Tên giá trị	*Số lượng*	*Tỷ lệ*
True	472	44.8%
False	581	55.2%

         Ví dụ: đối với thuộc tính danh nghĩa (giới tính)

Tên giá trị	*Số lượng*	*Tỷ lệ*
Male	472	42.9%
Female	581	52.8%
Khác	47	4.3%

         Ví dụ: đối với thuộc tính thứ tự (xếp loại)

Tên giá trị	*Số lượng*	*Tỷ lệ*
Xuất sắc	5	7.3%
Giỏi	12	17.6%
Khá	43	63.2%
Trung bình	6	8.8%
Yếu	2	3.1%

       - Đối với **tất cả các thuộc tính kiểu số**: Tính các giá trị mean, median, midrange, mode, min, max, five-number summary.
   - **Trình bày chi tiết** quá trình tiền xử lý dữ liệu với dữ liệu trên (nếu có).
2. **Phần 2: Phân tích – thống kê thủ công trên CSDL đã chọn**

   2.1. Tìm hiểu dữ liệu

   2.1.1. Chọn **tối thiểu 3 thuộc tính** để vẽ, mỗi thuộc tính cần vẽ các đồ thị sau:

- *Boxplot* dựa trên five-number summary.
- *Quantile–Quantile Plot* trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa.
- *Histogram* trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa.
- *Scatter* trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa.

2.1.2. Tương tự như phần 2.1.1, nhưng nhóm dữ liệu đang có theo 1 thuộc tính dạng danh nghĩa (ví dụ: loại hàng, tỉnh-thành phố, môn thi đấu thể thao, tên bệnh, …) Thực hiện vẽ các biểu đồ sau cho dữ liệu vừa nhóm:

- *Boxplot* dựa trên five-number summary
- *Histogram* trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

2.1.3. Đo lường sự tương đồng và khác biệt của dữ liệu bằng 2 cách: ma trận tương quan và độ đo Cosin

- Chọn tối thiểu 4 thuộc tính lần lượt thuộc các dạng dữ liệu sau:
  - Thuộc tính dạng danh nghĩa *(Nominal Attributes)*
  - Thuộc tính dạng nhị phân (*Binary Attributes)*
  - Thuộc tính dạng số (*Numeric Attributes)*
  - Thuộc tính dạng thứ tự (*Ordinal Attributes)*
- Chọn tối thiểu 4 dòng dữ liệu trong CSDL đã chọn trong phần 1(có thể chọn 4 dòng đầu tiên có trong dữ liệu để dễ quan sát).

Thực hiện so sánh kết quả của 2 cách đo lường sự tương đồng và khác biệt của dữ liệu.

2.2. Tiền xử lý dữ liệu: *(phải sử dụng hoàn toàn bằng mã lệnh Python và các thư viện/module của Python)*

2.3. Tổng hợp dữ liệu: *(phải sử dụng hoàn toàn bằng mã lệnh Python và các thư viện/module của Python)*

2.4. Trực quan hóa dữ liệu: *(phải sử dụng hoàn toàn bằng mã lệnh Python và các thư viện/module của Python)*

2.5. Thực hiện khai thác dữ liệu: *(phải sử dụng hoàn toàn bằng mã lệnh Python và các thư viện/module của Python)*

Sử dụng các phương pháp khai thác dữ liệu đã biết (tập phổ biến, phân lớp, phân cụm) để khai thác dữ liệu đã chọn trong phần 1. Với yêu cầu thực hiện tối thiểu 2 phương pháp bất kỳ do SV tự chọn (ví dụ sử dụng Apriori và FP-growth).

Yêu cầu thực hiện:

2.5.1. Trích 30% dữ liệu đang có (tạm gọi là tập D). Thực hiện tính toán thủ công 2 phương pháp khai thác dữ liệu đã chọn trên tập dữ liệu D này.

2.5.2. Với tập dữ liệu đầy đủ, cần thực hiện 2 việc:

(i).- Lập trình cho 2 phương pháp đã chọn

(ii).- Thực hiện đánh giá các mẫu thu được bằng các phương pháp đã biết bằng cách chọn 2 trong số các phương pháp đánh giá để đánh giá kết quả của việc thực hiện ở phần (i).

- *Nếu chọn tập phổ biến*: có thể chọn 2 trong các phương pháp sau:
  - Thang đo tương quan Lift
  - Thang đo χ^2
  - Độ đo tin cậy toàn phần (all_confidence)
  - Độ đo tin cậy tối đa (max_confidence)
  - Độ đo Kulczynski (Kulczynski measure)
  - Độ đo cosin (cosine measure)
- *Nếu chọn phương pháp phân loại*: có thể chọn 2 trong các phương pháp sau:
  - Số liệu để đánh giá hiệu suất phân loại *(Metrics for Evaluating Classifier Performance)*
  - Phương pháp Holdout và lấy mẫu con ngẫu nhiên *(Holdout Method and Random Subsampling)*
  - Xác thực chéo *(Cross-Validation)*
  - Phương pháp Bootstrap
  - Lựa chọn mô hình bằng cách sử dụng các thử nghiệm thống kê về tính quan trọng của mô hình *(Model Selection Using Statistical Tests of Significance)*
  - So sánh các phân loại dựa trên đường cong Chi phí-Lợi ích và ROC*(Comparing Classifiers Based on Cost–Benefit and ROC Curves)*
- *Nếu chọn phương pháp phân cụm*: sử dụng 2 phương pháp sau:
  - Phương pháp giám sát (*supervised methods*)
  - Phương pháp không giám sát (*unsupervised methods*)

PHỤ LỤC HƯỚNG DẪN TRÌNH BÀY BÁO CÁO

- **Một số quy định chung:**

  Khổ giấy A4. Lề Left=4cm, top=Bottom=Right=2cm

  Font name: Times New Roman, font size = 13

  Spacing before = 6 pt

  Line spacing = 1.2 pt

- **Bố cục báo cáo**

  Trang bìa

  Mục lục

  Nhận xét của Giảng viên

  Danh mục hình ảnh

  Danh mục bảng biểu

  Phần 1 của báo cáo

  Phần 2 của báo cáo

  Phần 3 của báo cáo

  Tài liệu tham khảo

📖 ***Nguồn dữ liệu phục vụ cho đề tài***: SV có thể tham khảo và download dữ liệu mẫu từ 1 trong các nguồn sau:

(i). Find Open Datasets and Machine Learning Projects | Kaggle

(ii). Data Sets (who.int)

(iii). DataBank | The World Bank

(iv). Data Center | Human Development Reports (undp.org)

(v). …