ĐỀ TÀI MÔN HỌC KHAI PHÁ DỮ LIỆU

1. MỤC TIÊU

- Nắm vững và vận dụng thành thạo quy trình Khai phá dữ liệu, từ tiền xử lý đến đánh giá mô hình.
- Áp dụng các thuật toán khai phá dữ liệu (phân cụm, phân lớp, luật kết hợp, …) để khám phá tri thức tiềm ẩn từ dữ liệu và rút ra insight có ý nghĩa.
- Thực hiện trọn vẹn pipeline Data Mining: tiền xử lý dữ liệu → mô hình hóa → đánh giá → diễn giải kết quả theo đúng chuẩn mực khoa học.
- Phát triển tư duy phân tích và đánh giá, biết nhận định ưu – nhược điểm của từng phương pháp, đồng thời đề xuất hướng cải tiến hoặc giải pháp tối ưu hơn.
- Khuyến khích sáng tạo và tư duy độc lập, thông qua việc tự **đặt câu hỏi nghiên cứu** và giải quyết vấn đề dựa trên bằng chứng dữ liệu.
- Có khả năng triển khai mô hình vào thực tế, thông qua xây dựng ứng dụng minh họa hoặc workflow tự động hóa (n8n, Streamlit, …).

2. HÌNH THỨC THỰC HIỆN

- Nhóm tối đa 4 sinh viên.
- Nếu làm theo nhóm: yêu cầu phân tích phải toàn diện, nhiều chiều hơn (số lượng câu hỏi, độ sâu phân tích, mô hình ML áp dụng).
- Có câu hỏi nghiên cứu? Thực hiện phân tích để làm rõ câu hỏi nghiên cứu.

3. NỘI DUNG YÊU CẦU

(1) Lựa chọn dữ liệu & mô tả dữ liệu

- Nguồn dữ liệu phải public, rõ ràng, có link.
- Mô tả dataset: số dòng, số cột, thuộc tính quan trọng.

(2) Tiền xử lý dữ liệu

- Làm sạch dữ liệu: xử lý thiếu, nhiễu, ngoại lệ.
- Chọn lọc thuộc tính (feature selection).
- Chuẩn hóa dữ liệu (nếu mô hình yêu cầu).

(3) Phân tích mô tả ban đầu

- Thống kê cơ bản.
- Biểu đồ: histogram, boxplot, scatter, heatmap correlation.
- Nhận xét và rút ra insight.

(4) Đề xuất các kỹ thuật Khai Phá Dữ Liệu

Các nhóm thuật toán sau:

A. Phân lớp (Classification): Naive Bayes, Decision Tree, Random Forest, SVM, …

B. Gom cụm (Clustering): K-Means, Hierarchical Clustering, …

C. Khai phá luật kết hợp: Apriori, FP-Growth

Có so sánh hiệu quả các mô hình.

(5) Đánh giá mô hình

- Classification → accuracy, confusion matrix, precision/recall, F1.
- Clustering → silhouette score, inertia.
- Association rules → support, confidence, lift.

(6) Trực quan hóa

- Minh họa kết quả bằng biểu đồ, biểu diễn cụm, cây quyết định, mạng luật kết hợp…

(7) Kết luận + hạn chế + hướng mở rộng
- Trả lời câu hỏi nghiên cứu.
- Rút ra insight.
- Đề xuất cải tiến.

(8) Nộp sản phẩm
- Báo cáo
- Slide trình bày
- Code minh chứng (Colab / Jupyter / Python script).

4. CẤU TRÚC BÁO CÁO

Chương 1: Giới thiệu
- Tổng quan đề tài
- Tình hình nghiên cứu
- Mục tiêu, câu hỏi nghiên cứu
- Phương pháp nghiên cứu, hướng tiếp cận của đề tài

Chương 2: Cơ sở lý thuyết
- Các khái niệm
- Thuật toán sử dụng
- Nghiên cứu liên quan

Chương 3: Dữ liệu & Phương pháp đề xuất
- Xác định vấn đề, tiền xử lý dữ liệu, …
- Phương pháp đề xuất

Chương 4 – Thực nghiệm & Kết quả & Thảo luận
- Thiết lập thực nghiệm, độ đo
- Các kết quả thí nghiệm
- Đánh giá và so sánh các mô hình

Chương 5 – Kết luận

5. QUY ĐỊNH TRUNG THỰC HỌC THUẬT
- Phải nộp sản phẩm gốc.
- Được phép thảo luận, nhưng phải ghi rõ nguồn hỗ trợ (nếu có).
- Nếu sử dụng code, script, notebook từ nguồn khác → cần trích dẫn rõ ràng.
- Nộp các code liên quan để minh chứng

6. YÊU CẦU DỮ LIỆU

Dữ liệu phải có nguồn gốc rõ rang (tổ chức tạo dữ liệu – uy tín, có bài báo trích dẫn); dữ liệu in-house khi dùng phải cẩn thận (dữ liệu sơ cấp như lấy thông tin trên mạng, giá cả - ổn), nhưng dữ liệu từ nguồn không uy tín, hay tự tạo thì phải cẩn thận.

RUBRIC ĐÁNH GIÁ

Tiêu chí	Mô tả yêu cầu	Điểm tối đa	Ghi chú
1. Tổng quan	- Chọn dataset rõ ràng, hợp lệ, có link nguồn. - Mô tả dữ liệu (kích thước, thuộc tính, ví dụ mẫu).	10 điểm	Nhận diện và lựa chọn nguồn dữ liệu phù hợp

	- Đặt mục tiêu/câu hỏi nghiên cứu cụ thể, khả thi.		
2. Tiền xử lý dữ liệu	- Làm sạch dữ liệu, xử lý missing values, chuẩn hóa định dạng. - Giải thích pipeline xử lý.	10 điểm	Thực hiện các bước tiền xử lý dữ liệu cho phân tích
3. Phân tích mô tả	- Thống kê cơ bản (mean, median, distribution, correlation). - Biểu đồ trực quan (histogram, scatter, boxplot, heatmap). - Giải thích insight từ EDA.	10 điểm	Trình bày kết quả phân tích dữ liệu mô tả
4. Mô hình	- Đề xuất thuật toán học máy: • Clustering (K-Means, DBSCAN, Hierarchical) hoặc • Classification (Logistic Regression, Decision Tree, Random Forest, …) * Luật kết hợp - Mô tả rõ thuật toán, cách chọn tham số. - Đánh giá kết quả (ví dụ: silhouette score, accuracy, confusion matrix).	30 điểm	Ứng dụng kỹ thuật phân tích nâng cao để rút ra tri thức
5. Trực quan hóa kết quả	- Biểu đồ minh họa cho phân tích mô tả & phân cụm/phân nhóm. - Biểu đồ rõ ràng, có chú thích, giải thích.	10 điểm	
6. Kết luận	- Trả lời câu hỏi/giả thuyết ban đầu. - Rút ra insight từ dữ liệu. - Nhận xét về hạn chế và hướng mở rộng.	5 điểm	Phân tích và đánh giá kết quả nghiên cứu dữ liệu
7. Trình bày báo cáo	- Viết đúng cấu trúc - Ngôn ngữ rõ ràng, logic, có mục lục, biểu đồ được gắn vào văn bản. - Ghi chú rõ ràng các tài liệu liên quan, tập tin ở phần cuối.	10 điểm	Trình bày kết quả phân tích dữ liệu bằng văn bản khoa học
8. Mã nguồn	- Nộp đầy đủ mã nguồn, có github ghi chú quá trình làm việc	15 điểm	
9. Điểm cộng	- Chương trình minh họa hay, gây ngạc nhiên - Nội dung nghiên cứu hay, gây ngạc nhiên	10 điểm	

---o0o--