Mini Project Nhóm – Nhập môn Khoa học Dữ liệu

1. Tên đề tài (gợi ý): Tự chọn đề tài

2. Mục tiêu:

Sinh viên áp dụng kiến thức đã học , chọn một dạng bài toán như Regression, Classification, Clustering, Anomaly Detection, NLP, Recommendation System, phân tích dữ liệu,... → **giải quyết một bài toán thực tế** do nhóm tự lựa chọn, đồng thời thực hành kỹ năng làm việc nhóm, tìm kiếm dữ liệu, xử lý và trực quan hóa thông tin.

---

3. Yêu cầu bài làm:

□ 1. Xác định bài toán

- Nhóm tự lựa chọn 1 chủ đề bất kỳ thuộc các lĩnh vực như:
    - Giáo dục, sức khỏe, tài chính, xã hội, mạng xã hội, hành vi tiêu dùng, v.v.
- Xác định **loại bài toán chính**:
  → Classification / Regression / Clustering / NLP / Anomaly Detection / Recommendation System
- Đặt ra **câu hỏi cụ thể hoặc mục tiêu**: Ví dụ: Dự đoán điểm thi, phát hiện giao dịch bất thường, gợi ý sản phẩm, phân nhóm khách hàng, phân loại cảm xúc,...

□ 2. Tự tìm kiếm và thu thập dữ liệu

- Tìm và chọn 1 bộ dữ liệu phù hợp từ các nguồn mở như Kaggle, UCI, Google Dataset Search, hoặc web scraping.
- Dữ liệu nên có ít nhất **500 dòng và ≥ 5 đặc trưng**.
- Mô tả ngắn gọn về nguồn và đặc điểm dữ liệu.

□ 3. Tiền xử lý và khám phá dữ liệu (EDA)

- Làm sạch dữ liệu (xử lý thiếu, ngoại lệ, mã hóa, chuẩn hóa…)
- Trực quan hóa dữ liệu để hiểu đặc điểm (sử dụng biểu đồ, phân phối, tương quan…)
- Nhận xét về những phát hiện đáng chú ý

□ 4. Xây dựng và huấn luyện mô hình

- Chọn một hoặc nhiều mô hình phù hợp với loại bài toán đã chọn (Linear Regression, Random Forest, SVM, KMeans, Logistic Regression, Naive Bayes, LSTM…)
- Chia dữ liệu thành train/test (hoặc cross-validation)
- Hiệu chỉnh siêu tham số nếu cần

□ 5. Đánh giá mô hình

- Sử dụng các chỉ số phù hợp: Accuracy, Precision, Recall, F1-score, RMSE, Silhouette score,...
- Đưa ra so sánh giữa các mô hình nếu có
- Nhận xét ưu/nhược điểm và tính ứng dụng thực tế

□ 6. Báo cáo và trình bày kết quả

- Báo cáo cần có cấu trúc rõ ràng (Làm trên file .ipynb)
    - Giới thiệu vấn đề
    - Dữ liệu & EDA
    - Xây dựng mô hình
    - Đánh giá & nhận xét
    - Kết luận & hướng mở rộng

Yêu cầu:

- Vận dụng được các nội dung đã học trên lớp, nắm bắt chu trình khoa học dữ liệu.
- Làm nhóm
- Thời gian: 3 tuần