YÊU CẦU CHUNG VỀ BÀI TẬP LỚN MÔN HỌC – K70
MÔN HỌC: KHAI PHÁ DỮ LIỆU

Mỗi nhóm 2 đến 3 sinh viên thực hiện các yêu cầu sau:

1. Tìm hiểu và mô tả về 01 bài toán ứng dụng phân lớp/phân cụm trong thực tế (như gợi ý bên dưới).
2. Thu thập và xử lý dữ liệu về bài toán đã lựa chọn.
3. Thực nghiệm thực nghiệm, phân tích và so sánh kết quả các phương pháp.
   - Nếu là bài toán phân lớp: thực nghiệm so sánh kết quả phân lớp bằng hai phương pháp Cây quyết định và Naive Bayes.
   - Nếu là bài toán phân cụm: thực nghiệm và so sánh kết quả phân cụm bằng hai phương pháp Kmean và Kmoid.

   Chú ý: Thực nghiệm trên python.
4. Trích khoảng 20 mẫu dữ liệu tự tập dữ liệu đã thực nghiệm của bài toán. Thực hiện tính toán mô phỏng cho hai thuật toán đã thực nghiệm ở câu 4.

Gợi ý một số bài toán phân lớp dữ liệu:

- Image classification
- Fraud detection
- Document classification
- Spam filtering
- Facial recognition
- Voice recognition
- Medical diagnostic test
- Customer behavior prediction
- Product categorization
- Malware classification
- Protein classification

Links:

https://vitalflux.com/classification-problems-real-world-examples/

https://research.aimultiple.com/image-classification/

Gợi ý các bài toán phân cụm dữ liệu:

https://datafloq.com/read/7-innovative-uses-of-clustering-algorithms/

https://www.tutorialspoint.com/what-are-the-applications-of-clustering

https://sites.google.com/site/dataclusteringalgorithms/clustering-algorithm-applications

https://link.springer.com/article/10.1007/s10462-022-10325-y

Dành cho các bạn học 2 môn KPDL và Bigdata:

- Community detection
- Node embedding
- Node classification pipeline
- Link prediction (dự đoán tương tác sinh học, liên kết mạng xã hội)

(https://graphacademy.neo4j.com/courses/graph-data-science-fundamentals/1-graph-algorithms/)