CHỦ ĐỀ ĐỒ ÁN MÔN HỌC KHAI KHOÁNG DỮ LIỆU

1. Xây dựng ứng dụng phân loại bản tin (ít nhất 5 chủ đề ví dụ xe, điện thoại, …):
   - Thu thập dữ liệu từ trang báo điện tử **https://genk.vn/**
   - Tiền xử lý dữ liệu.
   - Xây dựng mô hình phân loại (sử dụng ít nhất 3 giải thuật)
   - Đánh giá mô hình.
   - Xây dựng Restful API phân loại.
   - Xây dựng website/ứng dụng trên di động để phân loại bản tin mới.

2. Xây dựng ứng dụng nhận dạng các loại cá cảnh (ít nhất 10 loại):
   - Thu thập dữ liệu
   - Xây dựng mô hình (ít nhất 03 giải thuật)
   - Đánh giá mô hình.
   - Xây dựng Restful API.
   - Xây dựng website/ứng dụng trên di động để nhận dạng.

3. Phân tích thị trường lao động trên hệ thống Freelancer (Gợi ý thực hiện: cào dữ liệu về việc làm năm 2024 trên trang Freelencer. Phân tích tập dữ liệu này, ví dụ lương thấp nhất, cao nhất, công việc phổ biến nhất, ...). Liên hệ với giảng viên 2 tuần đầu tiên sau khi chọn đề tài để được cung cấp thêm tài liệu tham khảo.

4. Xây dựng website hỗ trợ sinh viên học gom nhóm dữ liệu với giải thuật cây phân cấp (hierarchical clustering). Website có các chức năng chính: upload tập tin dữ liệu có định dạng là csv hoặc excel, trình bảy các bước gom nhóm dữ liệu, chọn giải thuật xây dựng cây phân cấp (bottom-up hay top-down) hiển thị dữ liệu.

5. Phân tích tập dữ liệu bảo hiểm Insurance (https://github.com/ltdaovn/dataset/raw/master/insurance.xlsx). Các công việc chính:
   - Tìm hiểu dữ liệu
   - Tiền xử lý dữ liệu
   - Phân tích trực quan
   - Xây dựng mô hình máy học (hồi quy, phân lớp, gom nhóm, …)
   - Đánh giá mô hình
   - Xây dựng website để dự đoán số tiền bảo hiểm khách hàng mới có thể mua.

6. Phân tích tập dữ liệu Car Insurance 2024 (https://github.com/ltdaovn/dataset/raw/master/Car_Insurance2024.xlsx). Các công việc chính tương tự đề tài số 5.
7. Phân tích tập dữ liệu Glass2024 (https://github.com/ltdaovn/dataset/raw/master/glass.2024xlsx). Các công việc chính tương tự đề tài số 5.
8. Phân tích tập dữ liệu Dry Bean 2024 (https://github.com/ltdaovn/dataset/raw/master/Dry_Bean_Dataset2024.xlsx). Các công việc chính tương tự đề tài số 5.
9. Phân tích tập dữ liệu Maternal Health Risk 2024 (https://github.com/ltdaovn/dataset/raw/master/Maternal%20Health%20Risk%20Data%20Set_2024.xlsx). Các công việc chính tương tự đề tài số 5.
10. Phân tích tập dữ liệu Rice2024 (https://github.com/ltdaovn/dataset/raw/master/Rice2024.xlsx). Các công việc chính tương tự đề tài số 5.
11. Phân tích tập dữ liệu Breast Cancer (https://archive.ics.uci.edu/dataset/14/breast+cancer). Các công việc chính tương tự đề tài số 5.
12. Phân tích tập dữ liệu Nursery2024 (https://github.com/ltdaovn/dataset/raw/master/Nursery2024.xlsx). Các công việc chính tương tự đề tài số 5.
13. Phân tích tập dữ liệu Diabetes (https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008). Các công việc chính tương tự đề tài số 5.
14. Phân tích tập dữ liệu HCV (https://archive.ics.uci.edu/dataset/571/hcv+data). Các công việc chính tương tự đề tài số 5.
15. Phân tích tập dữ liệu Abalone 2024 (https://github.com/ltdaovn/dataset/raw/master/abalone2024.xlsx). Các công việc chính tương tự đề tài số 5.

16. Xây dựng mô hình dự báo tiền điện với tập dữ liệu the “European Wholesale Electricity Price” (https://ember-climate.org/data-catalogue/european-wholesale-electricity-price-data/).

17. Clustering tập dữ liệu hình các loại cây thuốc nam (https://drive.google.com/file/d/16IBw6t1i--xWMscB7bierxfg8lG5pMVM/view?usp=sharing)

18. Clustering tập dữ liệu Plants (https://archive.ics.uci.edu/dataset/180/plants)

19. Xây dựng mô hình nhận dạng tự động Tweets liên quan đến thảm họa. Dữ liệu học là các Tweets bằng tiếng Anh được gán nhãn là có liên quan đến thảm họa (ví dụ Damage to school bus on 80 in multi car crash) hoặc không có liên quan đến thảm họa (ví dụ My car is so fast). Yêu cầu: mô hình phân loại dữ liệu phải có độ chính xác lớn hơn 90%. Trong báo cáo, sinh viên cần mô tả từng chi tiết khi xây dựng mô hình. Dữ liệu tham khảo: https://raw.githubusercontent.com/ltdaovn/dataset/master/NLP%20with%20Disaster%20Tweets/Disaster_Tweets_train.csv

20. Xây dựng mô hình phân loại tự động các đánh giá về thuốc. Tương tự đề tài số 1 tuy nhiên học viên không cần thu thập dữ liệu mà sử dụng dữ liệu có sẵn tại địa chỉ https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com

21. Sinh viên có thể tự đề xuất đồ án. Tuy nhiên cần thảo luận trước với giảng viên.

Ghi chú chung:

- Nhóm tối đa 3 sinh viên (sinh viên có thể làm một mình).
- Đồ án chiếm 30% số điểm môn học.
    - Hoàn thành các yêu cầu chính của đề tài: 40%
    - Quyển báo cáo: 20%
    - Trình bày báo cáo: 20%
    - Báo cáo tiến độ hằng tuần đầy đủ: 10%
    - Chương trình viết dễ hiểu (đặt tên biến, tên hàm có ý nghĩa, ...): 10%
- Quyển báo cáo cần có:
    - Trang bìa ghi rõ tên đề tài và sinh viên thực hiện
    - Giới thiệu vấn đề
    - Giới thiệu tập dữ liệu
    - Cơ sở lý thuyết (trình bày các giải thuật được sử dụng trong đồ án)
    - Thiết kế và Cài đặt (tiền xử lý dữ liệu, phân tích trực quan, xây dựng mô hình, đánh giá mô hình).

    - Kết quả đạt được và hướng phát triển.
- Các sản phẩm cần nộp (tất cả nén lại thành 1 file zip). Tên file được đặt theo quy ước: NhomN.zip (N = 1, 2, …)
    - Code
    - Quyển báo cáo (file word và pdf)
- Sinh viên có thể sử dụng các thư viện có sẵn trong quá trình thực hiện đồ án.
- Các nhóm dùng github để quản lý code.
- Đối với các chủ đề xây dựng website hay webservices, sản phẩm cần được triển khai trên các đám mây miễn phí.