**DANH SÁCH GỢI Ý ĐỒ ÁN MÔN HỌC**

**MÔN: KHOA HỌC DỮ LIỆU ỨNG DỤNG**

1. **DANH SÁCH NHÓM ĐỀ TÀI**

|  |  |  |  |
| --- | --- | --- | --- |
| **TÊN ĐỀ TÀI** | **LOẠI ĐỀ TÀI** | **MÔ TẢ NỘI DUNG** | **GHI CHÚ** |
| **Gắn nhãn các đoạn tweet có nội dung phân biệt chủng tộc, phân biệt giới tính**  **(Twitter Sentimental Analysis)** | **Binary Classification/ text mining** | **Data:**  [**https://www.kaggle.com/datasets/cosmos98/twitter-and-reddit-sentimental-analysis-dataset?select=Twitter\_Data.csv**](https://www.kaggle.com/datasets/cosmos98/twitter-and-reddit-sentimental-analysis-dataset?select=Twitter_Data.csv)  **Yêu cầu:**   1. **Xử lý dữ liệu để tìm ra điểm chung của các tweet có nội dung phản cảm** 2. **Xây dựng mô hình phân loại nhị phân với nhãn:** 3. **Chạy các mô hình phân loại nhị phân đã học và đánh giá xem mô hình nào cho kết quả tốt hơn.** 4. **Xây dựng một bộ dữ liệu tweet thực tế và chạy thử để xem kết quả.** | **Để xử lý dữ liệu tốt, có thể import dữ liệu vào SQL, sau đó chạy các query để tìm ra các từ có nội dung phản cảm với tần suất lớn**  **Có thể tìm hiểu thêm các mô hình nhị phân bên ngoài và áp dụng (điểm cộng)**  **Xây dựng bộ tweet dựa trên việc thu thập các tweet thực tế hoặc sử dụng một phần bộ dữ liệu test** |
| **Định giá vé máy bay theo lịch trình và loại bay** | **Regression/ Prediction** | **Data:**  [**https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction?select=economy.csv**](https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction?select=economy.csv)  **Yêu cầu:**   1. **Tìm ra mối liên quan giữa thông số của các chuyến bay và giá tiền của chúng** 2. **Xây dựng mô hình định giá vé máy bay theo yêu cầu của người dùng** 3. **Chạy các mô hình máy học liên quan với bộ dữ liệu huấn luyện** 4. **Kiểm định mô hình với bộ dữ liệu thực tế** | **Để xử lý dữ liệu tốt cần tìm hiểu các yếu tố ảnh hưởng đến giá vé máy bay trong thực tế.**  **Những dữ liệu chưa được chuẩn hóa như thời gian cần được chỉnh lại (dùng excel hoặc sql)** |
| **Phân tích tình trạng nóng lên toàn cẩu thông qua số liệu tại các thành phố lớn trên thế giới** | **Clustering/ Classification** | **Data:**  **https://www.kaggle.com/datasets/hansukyang/temperature-history-of-1000-cities-1980-to-2020**  **Yêu cầu:**   1. **Tìm ra mối liên quan vị trí của các thành phố và nhiệt độ của chúng** 2. **Xây dựng mô hình dự báo nhiệt độ của một vị trí bất kỳ trên bản đồ vào một ngày cụ thể** 3. **Biểu diễn sự thay đổi nhiệt độ của các thành phố trong 40 năm dựa trên bản đồ nhiệt** | **Dữ liệu đang được biểu diễn theo dạng liệt kê, cần chuyển dạng dữ liệu** |