Lab 02: Khai thác và trực quan hóa dữ liệu

Võ Nhật Tân

1 Quy định chung

- Bài làm được thực hiện theo nhóm tối đa 3 người.
- Thành viên không tham gia sẽ không có điểm bài tập này.
- Bài làm phải tuân thủ theo yêu cầu đồ án.
- Các nguồn tài liệu tham khảo (nếu có) cần ghi đầy đủ trong báo cáo ở mục Tài liệu tham khảo.
- Các công cụ hỗ trợ trong việc code, viết báo cáo như ChatGPT, Github Copilot, ... chỉ dùng như một công cụ tham khảo; nội dung phải được kiểm tra và chỉnh sửa phù hợp với bài toán đặt ra. Nếu phát hiện sử dụng công cụ AI để sinh nội dung quá nhiều, hoặc nội dung không phù hợp/sai lệch thì sẽ trừ tối đa 50% số điểm (tùy mức độ).
- Bài giống nhau sẽ **0 điểm môn học**

2 Giới thiệu đồ án

2.1 Giới thiệu đồ án

World Development Indicators (WDI) là một nguồn dữ liệu do Ngân hàng Thế giới (World Bank) cung cấp, tập trung vào các chỉ số phát triển toàn cầu. Đây là kho dữ liệu toàn diện với hơn 1.600 chỉ số về kinh tế, xã hội và môi trường, bao gồm thông tin từ hơn 200 quốc gia trên toàn cầu, trải dài qua nhiều năm. Thông qua việc khai thác và phân tích dữ liệu này, đồ án hướng đến mục tiêu trình bày các xu hướng phát triển, khám phá mối quan hệ giữa các chỉ số, và làm rõ các vấn đề nổi bật như phát triển kinh tế, biến đổi khí hậu, ...

2.2 Nhiệm vụ đồ án

Trong đồ án này, các bạn sẽ thực hiện các nhiệm vụ sau:

1. Phân tích cơ bản về dữ liệu: giới thiệu dữ liệu, cỡ mẫu, cấu trúc, phân tích thống kê, xử lý dữ liệu, ...
2. Xác định mục tiêu và lựa chọn các trường dữ liệu
   - Xác định rõ mục tiêu và lựa chọn các trường dữ liệu liên quan để trực quan hóa
   - Cần xem xét các môi quan hệ giữa các trường để đưa ra lựa chọn phù hợp nhất.
   - Mỗi nhóm cần xác định ít nhất [**Số lượng thành viên**] * **3** mục tiêu để phân tích và trực quan.
3. Lựa chọn biểu đồ thích hợp và giải thích lý do
   - Biểu đồ cần phải phù hợp với tính chất của trường dữ liệu.
   - Nhóm sử dụng đa dạng các biểu đồ để làm rõ mục tiêu đã đề ra.
   - Toàn bài phân tích cần bao phủ hết các biểu đồ đã được học.
4. Nhận xét và đưa ra kết luận dựa trên các biểu đồ

3 Yêu cầu đồ án

- Nhóm thực hiện các nhiệm vụ trên trong môi trường lập trình Python, không sử dụng các phần mềm khác như Tableau, PowerBI, ... để minh họa; Các thư viện được sử dụng như *numpy, pandas, seaborn, matplotlib* được phép sử dụng; các thư viện ngoài thì cần ghi rõ trong báo cáo lý do sử dụng.
- Nhóm tổ chức tất cả các code Python trên một file Jupyter Notebook (**.ipynb**) và sử dụng các cell Markdown để mô tả rõ quá trình thực hiện.
- Nhóm chỉ sử dụng dữ liệu được cung cấp bởi World Development Indicators; không sử dụng các nguồn khác để so sánh, đối chiếu.
- Có thể sử dụng các thuật toán máy học trong việc phân tích và thể hiện trên biểu đồ.
- Viết báo cáo ngắn gọn và đầy đủ toàn bộ quá trình thực hiện; kết quả và nhận xét. Báo cáo tối đa là 24 trang (không bao gồm trang bìa, mục lục và tài liệu tham khảo). Nội dung file code và báo cáo phải khớp với nhau. Nội dung báo cáo bao gồm:
  - Thông tin nhóm: tên nhóm, mssv...
  - Mức độ hoàn thành tổng thể của mỗi yêu cầu.
  - Mức độ hoàn thành của từng thành viên.
  - Chi tiết các bước thực hiện (kèm hình ảnh), thuật toán, chạy ví dụ, nhận xét. Trình bày đơn giản, có hình minh họa.

4 Quy định nộp bài

- Nhóm sẽ cử đại diện 1 người nộp bài.
- Khi nộp bài, cần chạy lại các cell và không xóa output của các cell (đảm bảo biểu đồ đúng như báo cáo).
- Bài nộp là một file nén đặt tên là **[MSSV_1, MSSV_2, MSSV_3].zip** sẽ gồm:
  1. Thư mục chứa dữ liệu/Link tới dữ liệu
  2. File code: **[MSSV_1, MSSV_2, MSSV_3].ipynb**
  3. File báo cáo: **[MSSV_1, MSSV_2, MSSV_3].pdf**
- Trong trường hợp dữ liệu quá nặng, các bạn chỉ upload dữ liệu lên server ngoài như Google Drive, ..., nộp link và giữ link public ít nhất trong 2 năm; còn file code và báo cáo vẫn nộp trên Moodle.

5 Liên hệ

Mọi thắc mắc trong quá trình thực hiện vui lòng gửi mail về vntan.work@gmail.com