

TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN

CUỘC THI TRÍ TUỆ NHÂN TẠO
KHOA CÔNG NGHỆ THÔNG TIN NĂM 2022

ĐỀ BÀI VÒNG LOẠI

Thời gian: từ 29/06/2022 – 12/07/2022

I. Bài toán: Phân tích mối tương quan giữa tập dữ liệu dân số, GDP lực lượng lao động và tỷ lệ thất nghiệp.

Trong vòng loại này, chúng tôi cung cấp:

- Dữ liệu GDP của Việt Nam từ quý 1 năm 2005 đến quý 1 năm 2022.
- Dân số, tốc độ tăng trưởng dân số, lực lượng lao động, tỷ lệ thất nghiệp,..., từ năm 1990 đến năm 2021.

Các nhóm hãy thực hiện khai thác dữ liệu trên tập dữ liệu này, đồng thời thực hiện các yêu cầu như mô tả chi tiết trong Phần III dưới đây.

II. Bộ dữ liệu - Dataset

Bộ dữ liệu gồm 1 file: Vietnam-Macroeconomic-Data.xls, cụ thể như sau:

- Bộ dữ liệu GDP được mô tả chi tiết trong **Sheet GDP**, các dữ liệu chúng ta cần khai thác:

| Trường | Kiểu dữ liệu | Mô tả |
|----------------------|--------------|--|
| Quý và tổng số tháng | String | Quý hiện tại và tổng của dữ liệu |
| GDP tổng | Float | Tổng sản phẩm nội địa, tức tổng sản phẩm quốc nội hay GDP (Gross Domestic Product) là giá trị thị trường của tất cả hàng hóa và dịch vụ cuối cùng được sản xuất ra trong phạm vi quốc gia Việt Nam trong thời kỳ một năm (đơn vị tính: tỉ đồng và tỷ lệ phần trăm) |

| | | |
|-------------------------|-------|---|
| Chi tiết GDP từng ngành | Float | Chi tiết GDP từng ngành (đơn vị tính: tỉ đồng và tỷ lệ phần trăm) |
|-------------------------|-------|---|

Bảng 1: Bộ dữ liệu GDP

- Bộ dữ liệu dân số được mô tả chi tiết trong **Sheet Dân_số_VN** từ năm 1990 đến năm 2021, các dữ liệu chúng ta cần khai thác sau đây:

| Trường | Kiểu dữ liệu | Mô tả |
|--------------------|--------------|--|
| Năm | Integer | Quý hiện tại và tổng của dữ liệu |
| Dân số | Float | Tổng dân số Việt Nam (ĐTV: triệu dân) |
| Tỷ lệ tăng dân số | Float | Tỷ lệ tăng dân số qua các năm (ĐTV: tỷ lệ phần trăm) |
| Lực lượng lao động | Float | Số lượng lao động qua các năm (ĐTV: triệu dân) |
| Thất nghiệp | Float | Số lượng người từ nguồn lao động sẵn có không tìm được việc làm (ĐTV: triệu dân) |

Bảng 2: Bộ dữ liệu Dân số

III. Yêu cầu

Toàn bộ quy trình bao gồm nhiều bước từ tiền xử lý đến đánh giá. Các nhóm được yêu cầu khai phá dữ liệu và báo cáo công việc của mình từng bước sau đây:

1. Xử lý dữ liệu - Data Imputation (4 điểm):

Có một số giá trị bị thiếu trong tập dữ liệu đã cho, các nhóm có thể đề xuất cách điền vào những giá trị còn thiếu đó không?

Lựa chọn những dữ liệu đáp ứng yêu cầu.

2. Khám phá dữ liệu- Data Exploration (6 điểm):

Nhóm cần khám phá dữ liệu để hiển thị một số thông tin thống kê và phân tích của tập dữ liệu đã lựa chọn.

Hướng dẫn: Các nhóm có thể thống kê về GDP và tỷ lệ thất nghiệp sau đó trực quan các kết quả này; có mối quan hệ nào giữa GDP và tỷ lệ thất nghiệp hay không. Sự thấu hiểu thông tin thông qua việc trực quan, hiển thị thông tin thống kê về dữ liệu sẽ hữu ích cho các nhóm trong các bước tiếp theo.

3. Trích xuất đặc trưng- Feature Extraction (4 điểm):

Hãy đề xuất cách trích xuất đặc trưng từ bộ dữ liệu đã cho, cung cấp lý do và giải thích cách làm của các nhóm.

Hướng dẫn:

- Các nhóm có thể sử dụng các biểu đồ để biểu diễn các đặc trưng của dữ liệu.
- Phương pháp Ranking.
- Biểu diễn sự phân bố dữ liệu.
- ...

4. Dự đoán- Prediction (5 điểm):

Hãy dự đoán tỉ lệ thất nghiệp, GDP theo các đặc trưng mà nhóm trích xuất ở trên và đánh giá trên bộ dữ liệu đã cho với tỉ lệ Train/Test là 8:2 với các độ đo phù hợp.

5. Thảo luận (1 điểm):

Đây là một nhiệm vụ của ngành khoa học dữ liệu và Trí tuệ nhân tạo khi giải quyết những thử thách trong nhu cầu thực tế hiện nay, các nhóm có thể ra ý kiến, giải pháp của mình những vấn đề liên quan. Các nhóm vui lòng cung cấp các thảo luận, ý tưởng của nhóm về việc thu thập các đặc trưng mới ột cách hiệu quả để tăng hiệu suất của mô hình.

IV. Báo cáo

- Báo cáo của các nhóm phải được sử dụng ngôn ngữ tiếng Việt.
- Báo cáo phải có cấu trúc rõ ràng cho từng yêu cầu ở trên, với mỗi yêu cầu đều bao gồm: phần giới thiệu, cách tiếp cận, đánh giá, thảo luận, v.v. phù hợp để các nhóm hoàn thành các yêu cầu ở trên một cách hiệu quả (*theo mẫu đính kèm trong thông báo*).

V. Nộp bài

Vui lòng gửi báo cáo của nhóm (ở định dạng pdf) và mã (bằng Python) tại địa chỉ <http://hackathon.daihocsonnt.edu.vn/> trước **24h00 ngày 12/07/2022**.

TM. BAN TỔ CHỨC CUỘC THI
TRƯỞNG BAN TỔ CHỨC



TS. Nguyễn Kim Quốc