CHƯƠNG TRÌNH TRÍ TUỆ NHÂN TẠO VÀ HỆ THỐNG THÔNG TIN
ĐOÀN THANH NIÊN VIỆN KỸ THUẬT CÔNG NGHỆ

CUỘC THI
KHAI PHÁ DỮ LIỆU
(TDMU- Entropy Data Analytics HACKATHON, Lần I-2021)

ĐỀ BÀI VÒNG LOẠI

Thời gian: từ 5/04/2021 – 12/4/2021

I. Bài toán: Phân tích kinh tế vĩ mô Việt Nam

Trong vòng này, chúng tôi cung cấp dữ liệu kinh tế vĩ mô của Việt Nam. Bạn sẽ thực hiện khai phá dữ liệu trên tập dữ liệu này và 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
với 3 hàng và 40 cột
Bộ dữ liệu được mô tả chi tiết trong Bảng 1 sau đây:

Trường	**Kiểu dữ liệu**	**Mô tả**
Year	integer	Năm hiện tại của dữ liệu
GDP	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 đô).
Unemployment rate	float	Phần trăm số người từ nguồn lao động sẵn có không tìm được việc làm

Bảng 1: Mô tả về bộ dữ liệu Vietnam-Macroeconomic-Data.xls

III. Yêu cầu

Toàn bộ quy trình bao gồm nhiều bước từ tiền xử lý đến đánh giá. Bạn 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 (20 điểm): Có một số giá trị bị thiếu trong tập dữ liệu đã cho, bạn có thể đề xuất cách điền vào những giá trị còn thiếu đó không?

2. Khám phá dữ liệu- Data Exploration (30 điểm): Bạn cần khám phá dữ liệu để hiển thị một số thông tin thống kê và phân tích của tập dữ liệu đã cho. Chẳng hạn như thống kê về GDP và Tỷ lê thất nghiệp; 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 sẽ hữu ích cho bạn trong các bước tiếp theo.

3. Trích xuất đặc trưng- Feature Extraction (15 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 bạn.

4. Dự đoán- Prediction (20 điểm): Hãy dự đoán tỉ lệ thất nghiệp theo các đặc trưng bạn 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 (15 điểm): Đây là một nhiệm vụ của khoa học dữ liệu khi giải quyết một thử thách, bạn được yêu cầu đưa ra ý kiến của mình về các giải pháp hiện tại của bạn cho thử thách này. Bạn cũng vui lòng cung cấp các thảo luận, ý tưởng của bạn về thu thập các đặc trưng mới, hiệu quả để tăng tăng hiệu suất của mô hình.

IV. Báo cáo

Báo cáo của bạn được trình bày bằng hình thức tốt, 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 bạn hoàn thành các yêu cầu ở trên một cách hiệu quả.

V. Nộp bài

Vui lòng gửi báo cáo của bạn (ở định dạng pdf) và mã (bằng Python) về địa chỉ email **hung.buithanhcs@gmail.com** trước **24h ngày 12/04/2021**.

BAN TỔ CHỨC TDMU- Entropy Data Analytics HACKATHON, 2021