CSC12110 – Phân tích dữ liệu ứng dụng

ĐỒ ÁN THỰC HÀNH

I. Thông tin chung

Mã số bài tập:	**DATH#2**
Thời lượng dự kiến:	10 tuần
Deadline nộp bài:	
Hình thức:	Bài tập cá nhân
Hình thức nộp bài:	Nộp qua Moodle môn học
GV phụ trách:	Hồ Thị Hoàng Vy
Thông tin liên lạc với GV:	hthvy@fit.hcmus.edu.vn,

II. Chuẩn đầu ra cần đạt

Bài tập này nhằm mục tiêu đạt được các chuẩn đầu ra sau:

- G1.1 - Thành lập, tổ chức, vận hành và quản lý nhóm
- G1.3 - Phân tích, tổng hợp và viết tài liệu kỹ thuật theo mẫu cho trước theo cá nhân hoặc cộng tác nhóm
- G6.1 - Hiểu và áp dụng được các phương pháp hồi quy, máy học
- G7.1 - Có kỹ năng làm việc với ngôn ngữ Python, cụ thể với Pandas, Numpy
- G7.2 - Có kỹ năng làm việc với các thư viện của Python : Statsmodels, Scikit-learn
- G7.3 - Có kỹ năng sử dụng thư viện Matplotlib, Seaborn để trực quan hóa dữ liệu

III. Mô tả bài tập

Dataset: Adventurework 2012
Sử dụng file **.bak** để khôi phục cơ sở dữ liệu mẫu vào phiên bản SQL Server của bạn.
Adventure Works Cycles là một công ty sản xuất lớn, đa quốc gia chuyên sản xuất và phân phối xe đạp bằng kim loại và composite cho các thị trường thương mại ở Bắc Mỹ, Châu Âu và Châu Á. Trụ sở chính của Adventure Works Cycles là Bothell, Washington, quy mô 500 công nhân. Ngoài ra, Adventure Works Cycles còn có một số đội bán hàng trên toàn cơ sở thị trường của mình thuộc các khu vực (region).
Adventure Works Cycles hiện muốn mở rộng thị phần của mình bằng cách nhắm mục tiêu quảng cáo đến những khách hàng tốt nhất của mình, mở rộng tính khả dụng của sản phẩm thông qua trang Web bên ngoài và giảm chi phí bán hàng bằng cách giảm chi phí sản xuất
Yêu cầu:

1. Dự đoán doanh số sản phẩm bán ra cho năm tiếp theo
2. Dự đoán 1 khách hàng sẽ mua sản phẩm X hay không?
3. Sử dụng thuật toán gom cụm để phân cụm khách hàng

4. Sinh viên tự xác định 1 trường hợp phân tích vận dụng (1 trong những thuật toán đã học) để phân tích cho phần sản xuất và nhân sự trong cơ sở dữ liệu.

IV. Các yêu cầu & quy định chi tiết cho bài nộp

- Phân tích yêu cầu, xác định và rút trích dữ liệu liên quan đáp ứng các nhu cầu trên
- Vận dụng quy trình phân tích dữ liệu trên lý thuyết (Data Analysis Process) để giải quyết các yêu cầu trên
- Đánh giá mô hình
- Gợi ý
    - Tìm hiểu và mô tả các biến cần cho mỗi yêu cầu.
    - Thống kê, vẽ biểu đồ để thể hiện phân phối của dữ liệu, đánh giá sơ bộ về tính chất của các biến.
    - Xử lý dữ liệu (thừa, thiếu, outlier…)
    - Xác định outcome hay các biến cần dự đoán
    - Loại dữ liệu đang có và loại thông tin trong mỗi cột (dependent variables, explanatory variables)?
    - Xây dựng mô hình, đánh giá mô hình
    - Nhận xét kết quả.

V. Cách đánh giá

a. Report document:
    i. Thông tin thành viên nhóm, phân công công việc, kết quả đạt được
    ii. Đánh giá kết quả đạt được của mỗi thành viên
    iii. Báo cáo chủ yếu mô tả lại toàn bộ quy trình từ giai đoạn phân tích yêu cầu đến giai đoạn hiển thị và nhận xét kết quả thực hiện.
b. Source code thực thi kèm các file dữ liệu liên quan