

CSC12110 – Phân tích dữ liệu ứng dụng

ĐỒ ÁN THỰC HÀNH #1

I. Thông tin chung

Mã số bài tập:	2425.DATH#1
Thời lượng dự kiến:	8 tuần
Deadline nộp bài:	Xem moodle
Hình thức:	Nhóm
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

III. Mô tả bài tập

- Dataset: xem Moodle

Yêu cầu:

- **Áp dụng các kiến thức trên lý thuyết cho các yêu cầu liên quan.**
- **Quan sát và phân tích tập dữ liệu (EDA), tiền xử lý dữ liệu và cài đặt mô hình với 1 số yêu cầu gợi ý sau:**
 1. Kiểm tra dữ liệu missing, duplicates, outliers
 2. Kiểm tra sự cân bằng giữa nhãn: click/không click quảng cáo
 3. Xử lý dữ liệu missing, duplicates, outliers
 4. Phân tích đơn biến (Univariate Analysis), Bivariate Analysis
 - VD: Quan sát tỷ lệ click quảng cáo theo từng độ tuổi; theo thu nhập, giới tính. Nhận xét
 - VD: Kiểm tra sự tương quan giữa các thuộc tính đối với việc nhấp quảng cáo.
 - Loại bỏ những thuộc tính không cần thiết
 5. Kết luận các điểm chính quan sát được từ dữ liệu từ các gợi ý trên
 6. Quan sát và nhận xét về:
 - Thời gian sử dụng website theo độ tuổi, theo thu nhập hoặc theo thành phố sinh sống.
 - Thời điểm trong ngày (sáng, trưa, chiều, tối)
 - Chủ đề quảng cáo được quan tâm nhất hoặc ít được quan tâm nhất
 7. Dự đoán khả năng 1 người dùng có chọn xem quảng cáo được hiển thị hay không (lựa chọn và cài đặt thuật toán, cài đặt ít nhất 2 thuật toán để so sánh với nhau)
 - Kiểm tra sự ảnh hưởng của thuộc tính thu nhập có ảnh hưởng đến mô hình dự đoán trên hay không, giải thích và kết luận.
 - Các thuộc tính nào có mối tương quan đến dự đoán khả năng click/không vào quảng cáo?

8. Đánh giá chất lượng mô hình sử dụng cross validation với các độ đo precision, recall, f1 trên tập train, test. Kết luận mô hình được lựa chọn

IV. Các quy định khác

- Áp dụng các giai đoạn được học trên lớp lý thuyết vào đồ án.
- Nộp lên moodle file .ipynb và link colab có share quyền chỉnh sửa cho giáo viên
- Đánh giá tỷ lệ tham gia của từng thành viên
- Các nhận xét, kết luận cần chi tiết, đầy đủ