**Đề Bài Tập Buổi 4 và 5**

Lưu ý: Các file dữ liệu được đính kèm theo tập tin này trên Google Classroom

**Câu 1**

Cho bộ dữ liệu gồm file mushroom.csv, cột “class” chứa thông tin cần phân loại: nhóm của loại nấm.

Yêu cầu:

1. Tiền xử lý dữ liệu
2. Chia bộ dữ liệu theo tỉ lệ train/test là 70/30
3. Tiến hành xây dựng mô hình Decision Tree cho bộ dữ liệu
4. (Câu hỏi nâng cao) bạn có nhận xét gì về performance của model? Đề xuất và triển khai thử 1 cách để cải thiện performance của model
5. Visualize cây quyết định ở câu (3) bằng GraphViz

**Câu 2**

Cho bộ dữ liệu gồm file Bank.csv, cột “y” chứa target variable là khách hàng có quyết định gửi tiền hay không.

Yêu cầu:

1. Tiền xử lý dữ liệu
2. Chia bộ dữ liệu theo tỉ lệ train/test là 70/30
3. Tiến hành xây dựng mô hình thích hợp cho bộ dữ liệu
4. Đánh giá performance của model
5. Thử sử dụng SMOTE và nhận xét hiệu quả của phương pháp này đối với performance của model
6. Thử sử dụng class weight và nhận xét hiệu quả của phương pháp này đối với performance của model

**Câu 3**

Cho bộ dữ liệu gồm file titanic\_csv.csv, cột “survived” chứa target variable là liệu hành khách có sống sót được trong thảm họa không.

Yêu cầu:

1. Tiền xử lý dữ liệu
2. Chia bộ dữ liệu theo tỉ lệ train/test là 70/30
3. Tiến hành xây dựng mô hình thích hợp cho bộ dữ liệu
4. Đánh giá performance của model

Lưu ý: mỗi câu được làm riêng ra 1 file jupyter notebook, các file notebook nên được nén thành 1 tập tin zip trước khi upload lên classroom