**Bài Tập Buổi 3**

Lưu ý: Bộ dữ liệu đi kèm để làm các bài tập được đính kèm chung với file này trong classroom

**Exercise 1: Multiclass Classification with KNN**

Cho bộ dữ liệu glass.data.txt

Yêu cầu:

1. Thực hiện việc chia bộ dữ liệu theo tỉ lệ 70-30 (train/test), với random seed được set cố định
2. Cho biết cột cuối cùng chứa nhãn cần phân loại (label), bạn có nhận xét gì về phân phối các nhãn trong cột này?
3. Nêu các bước tiền xử lý dữ liệu, và thực hiện tiền xử lý dữ liệu
4. Train một mô hình kNN với k=5, print classification\_report trên tập test, từ đó nhận xét hiệu quả của mô hình hiện tại?
5. Chọn giá trị k tối ưu cho mô hình kNN trên mô hình hiện tại (kèm 1 dòng giải thích tại sao lại chọn giá trị k này)
6. Câu hỏi nâng cao: Theo bạn kNN có thể dùng để phân loại văn bản không? Nếu có thể, bạn sẽ xử lý dữ liệu dạng text như thế nào?

**Exercise 2: Customer Spend Regression with KNN**

Cho bộ dữ liệu wrangled\_transactions.txt

Yêu cầu:

1. Cho biết cột mang giá trị cần dự đoán là “2011 revenue” - the last column, hãy đọc bộ dữ liệu này và chia theo tỉ lệ 70-30 (train/test) với random set cố định
2. Tiến hành tiền xử lý dữ liệu
3. Train 1 mô hình kNN để dự đoán giá trị trong cột “2011 revenue”, với giá trị k=5
4. Tiến hành tính Mean Square Error trên tập test, và in giá trị Mean Square Error tại k=5
5. Tiến hành việc chọn k thông qua việc khảo sát Mean Square Error, vẽ 1 đồ thị để biểu thị khi thay đổi k thì Mean Square Error trên 2 tập train và test thay đổi như thế nào?

**Note:** Với mỗi bài tập, học viên làm ra 1 file jupyter notebook và nén lại để submit trên Google Classroom.

Cách submit tương tự như ở bài tập buổi 2