**XÂY DỰNG MÔ HÌNH DỰ BÁO TÌNH TRẠNG CỦA CÁC KHOẢN NỢ**

Sử dụng bộ dữ liệu “data\_credit.csv” về tình trạng các khoản nợ của khách hàng tại ngân hàng Thera. Dữ liệu bao gồm các biến:

Biến mục tiêu (Target Variable) là “Loan Status” bao gồm 2 trạng thái

“Fully Paid” và “Charged Off”.

Các biến đầu vào còn lại trong bộ dữ liệu bao gồm:

* mã khách hàng (Customer ID),
* mã khoản vay (Loan ID),
* **số tiền cho vay hiện tại (Current Loan Amount),**
* kỳ hạn vay (Term),
* tình trạng sở hữu nhà (Home Ownership),
* mục đích vay (Purpose),
* **số tiền nợ thanh toán hàng tháng (Monthly Debt),**
* **số năm vay nợ (Years of Credit History),**
* **số lượng tài khoản đã mở (Number of Open Accounts),**
* số lần phát sinh vấn đề liên quan đến các khoản nợ (Number of Credit Problems),
* **số dư nợ hiện tại (Current Credit Balance).**

Yêu cầu:

1. Tiến hành khai báo dữ liệu. Cho biết dạng dữ liệu của các biến trong tập dữ liệu? Cho biết thông tin về các biến bị thiếu dữ liệu?

2. Thống kê mô tả các biến trong bộ dữ liệu.

3. Vẽ đồ thị cho biết tình trạng của các khoản nợ trong bộ dữ liệu.

4. Xác định 5 biến có tương quan cao nhất với biến mục tiêu để xây dựng mô hình dự báo tình trạng của các khoản nợ.

5. Xây dựng mô hình dự báo tình trạng của các khoản nợ với 5 biến đầu vào xác định từ câu 4. Chia tập dữ liệu theo các tỷ lệ giữa tập train và tập test như sau: 80:20, 70:30, 60:40. Với từng tập dữ liệu được chia, thực hiện việc lựa chọn mô hình tốt nhất theo các thuật toán:

a. Logistic Regression

b. KNN

c. Decision tree

d. Random forest

Việc lựa chọn được dựa trên tính toán Accuracy\_score và ma trận nhầm lẫn (Confusion Matrix).