ĐỒ ÁN CUỐI KỲ – MÔN KHOA HỌC DỮ LIỆU

Đề tài: Phát hiện gian lận thẻ tín dụng (Credit Card Fraud Detection)

1. Thành viên nhóm

Nguyễn Thành Nam – N21DCCN150
Sỳ Hưng – N22DCCN137
Uông Ngọc Sơn – N22DCCI033

2. Mục tiêu đề tài

Xây dựng và đánh giá mô hình nhằm phát hiện các giao dịch gian lận thẻ tín dụng trên tập dữ liệu có mức độ mất cân bằng cao.

3. Dữ liệu đầu vào

Tập dữ liệu sử dụng: creditcard.csv (gồm các đặc trưng V1-V28 đã được ẩn danh, cùng Time, Amount và nhãn Class). Bài toán có tính chất mất cân bằng mạnh: số giao dịch gian lận chiếm tỉ lệ rất nhỏ so với giao dịch bình thường.

4. Quy trình thực hiện

- Nạp và mô tả dữ liệu
- Phân tích khám phá dữ liệu (EDA)
- Tiền xử lý và chia tập train/test
- Huấn luyện mô hình Logistic Regression và Random Forest
- Đánh giá bằng ROC-AUC, PR-AUC, Precision, Recall, F1-score
- Tối ưu threshold và cross-validation

5. Phân công nhiệm vụ

Nguyễn Thành Nam: Nạp dữ liệu, kiểm tra chất lượng dữ liệu, EDA và trực quan hóa
Sỳ Hưng: Tiền xử lý dữ liệu, xây dựng pipeline và huấn luyện mô hình
Uông Ngọc Sơn: Đánh giá mô hình, threshold tuning, cross-validation, tổng hợp báo cáo

6. Kết quả chính

Các mô hình cho thấy khả năng phát hiện gian lận tốt hơn so với baseline. Logistic Regression cho Recall cao trong khi Random Forest cho Precision và PR-AUC tốt hơn.

7. Github demo

Link Github: https://github.com/ThanhNam-42/CreditCardFraudDetection