ARTI 308 - Machine Learning

Lab 2: Identifying ML Problems, Selecting Open Datasets, and Drawing a Methodology Diagram

---

SECTION 1: DATASET DESCRIPTION

Dataset Name: Credit Card Fraud Detection
Source: Kaggle - ULB Machine Learning Group
Format: CSV
Size: 150 MB
Rows: 284,807 transactions
Columns: 31 features

Features:

Time, V1-V28 (PCA transformed features), Amount, Class

Target Variable: Class

0 = Legitimate transaction (284,315 cases)

1 = Fraudulent transaction (492 cases)

Class Imbalance: Only 0.17% of transactions are fraudulent.

SECTION 2: MACHINE LEARNING PROBLEM DEFINITION

1. What type of ML problem is this?

This is a **binary classification** problem. We are trying to predict one of two outcomes: either a transaction is fraud (1) or it's not fraud (0). This is not regression because we are not predicting a continuous number, and it's not clustering because the dataset already has labels.

---

2. Is there a target variable?

Yes. The target variable is called Class.

Class = 0 , Legitimate transaction

Class = 1 , Fraudulent transaction

---

3. What will the model learn/predict?

The model will learn patterns from past transactions (like transaction amount, time, and the V1-V28 features) to predict whether a **new transaction** is fraudulent or legitimate. The model will output either:

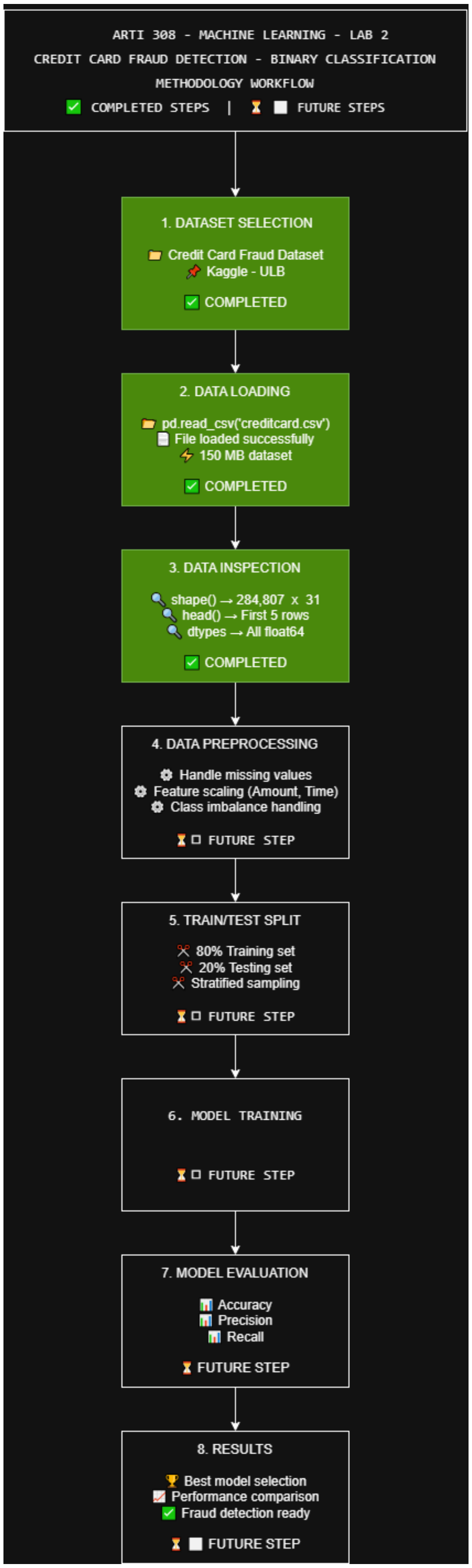
0 - Legitimate transaction (safe to process)

1 - Fraudulent transaction (flag for review)

Summary :

We are using the Credit Card Fraud Detection dataset from Kaggle. It has 284,807 transactions and 31 columns. Our goal is to build a machine learning model that can automatically detect fraud. The hardest part is that only 492 out of 284,807 transactions are fraud (0.17%), so the data is very imbalanced.

Methodology Diagram:



ARTI 308 - MACHINE LEARNING - LAB 2
CREDIT CARD FRAUD DETECTION - BINARY CLASSIFICATION
METHODOLOGY WORKFLOW
✓ COMPLETED STEPS | ⌛ □ FUTURE STEPS

1. DATASET SELECTION
Credit Card Fraud Dataset
Kaggle - ULB
✓ COMPLETED

2. DATA LOADING
pd.read_csv('creditcard.csv')
File loaded successfully
150 MB dataset
✓ COMPLETED

3. DATA INSPECTION
shape() → 284,807 x 31
head() → First 5 rows
dtypes → All float64
✓ COMPLETED

4. DATA PREPROCESSING
Handle missing values
Feature scaling (Amount, Time)
Class imbalance handling
⌛ □ FUTURE STEP

5. TRAIN/TEST SPLIT
80% Training set
20% Testing set
Stratified sampling
⌛ □ FUTURE STEP

6. MODEL TRAINING
⌛ □ FUTURE STEP

7. MODEL EVALUATION
Accuracy
Precision
Recall
⌛ FUTURE STEP

8. RESULTS
Best model selection
Performance comparison
✓ Fraud detection ready
⌛ □ FUTURE STEP

The link (For Better View): Methodology Diagram Link