# Lab1_KNN_Classifier Report

**a)** What is fraud detection, and why is it important?

Fraud detection is the process of identifying fraudulent activities — such as unauthorized transactions, identity theft, insurance scams, etc. — using data and statistical or machine learning techniques.

As per the dataset Class column: 0 = **non-fraud** and 1 = **fraud**

**Why is it important?**

- Prevent **financial losses**
- Protect **user trust and brand reputation**
- Comply with **regulatory requirements**

**b)** If you change the training and testing split to 70% training and 30% testing, how does the model's performance change?

When I change training data from 60% to 70%, the model performance on the test data shows:

- Slight improvement in **ROC AUC** (0.9186 → 0.9287), meaning the model **discriminates better** between fraud and non-fraud and has a **better balance between TPR and FPR.**
- **Precision** stays high (0.97), indicating good positive predictions.
- Slight drop in **recall and F1-score**, likely because of **fewer fraud cases** in the test set (137 → 107).

| Metric | 40% Test (60% Training) | 30% Test (70% Training) | |
|---|---|---|---|
| ROC AUC | 0.9186 | 0.9287 | Slightly better (high) |
| Precision | 0.97 | 0.97 | same |
| Recall | 0.80 | 0.79 | Slightly drop |
| F1-Score | 0.88 | 0.87 | Slightly drop |
| Accuracy | 1.00 | 1.00 | same |

**c)** Keeping the test size fixed at 40%, try changing the number of neighbors (in **KNN**). How does the model's performance vary with different **K** values?
Which value gives the best result, and how do you define what makes it the "**best**"?

As per the result, K= 1 to 11, the best result gave K=3 or 5, because it has the best **F1-Score, Recall** for fraud class(**better balance between precision & recall)**. I will choose K=5 it has slightly higher AUC then K=3.

| K Value | Precision(fraud) | Recall(fraud) | F1-Score(fraud) | ROC AUC(fraud) |
|---|---|---|---|---|
| 1 | 0.84 | 0.82 | 0.83 | 0.9075 |
| 3 | 0.97 | 0.80 | 0.88 | 0.9153 |
| **5** | **0.97** | **0.80** | **0.88** | **0.9186** |
| 7 | 0.97 | 0.79 | 0.87 | 0.9256 |
| 9 | 0.97 | 0.77 | 0.86 | 0.9252 |
| 11 | 0.97 | 0.77 | 0.86 | 0.9249 |

Precision is lower at K=1 (0.84) __ more false positives at this point, but K=3 to 11 stays high nearly (0.97)
Recall is highest at K= 1 (0.82), decreases slightly when k in increases. This means that model detect frauds less likely as K increases.
F1-Score(Balance of Precision and Recall) is lower at K=1(0.83) due to low precision, peaks at K=3 and K=5 (0.88) and slightly drops for K= 7(0.87), K=9, K=10 (0.86)
ROC AUC Score improves from K=1 (0.90) to K=7 (0.92) and slightly drops at K=9 and K=10, best AUC is K=7.