# Practical Class Work - 3

## Title: Loan Approval Risk Model

### Business Goal

A bank wants to reduce losses. For each applicant, we must predict:

- **Bad risk (1)** vs **Good risk (0)**

and produce a **simple business decision recommendation**.

---

## Notebook Structure and Task Flow

### 0) Notebook Setup, required libraries:

- pandas, numpy
- matplotlib
- scikit-learn

---

## 1) Data Loading and Business Understanding

### Task 1.1 Load dataset

- Load German Credit dataset into pandas DataFrame.

  https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

- Show:
  - number of rows/columns
  - first 5 rows
  - target distribution (how many good vs bad)

### Task 1.2 — Business interpretation

Write 5–7 lines (markdown) explaining:

- What is a “default / bad risk” in a bank context?
- Why mistakes matter (false negatives vs false positives)

# 2) Data Quality and Simple EDA

### Task 2.1 Missing values & data types

- Check missing values
- Identify numeric vs categorical features

### Task 2.2 Correlation

- Compute correlation matrix for numeric columns
- Pick **Top-3 numeric features** most correlated with the target

---

# 3) Preprocessing

### Task 3.1 Split train/test

- Split into train/test (e.g., 80/20)
- Use random_state for reproducibility

### Task 3.2 Prepare pipeline

- Categorical → OneHotEncode
- Numeric → StandardScaler (if required)

---

# 4) Build 3 Baseline Models (No Tuning)

Train **exactly these 3** models:

1. **Logistic Regression**
2. **Decision Tree**
3. **kNN**

### Task 4.1 Train models

- Fit each model using the same preprocessing pipeline.

### Task 4.2  Evaluate models

For each model, compute:

- Accuracy
- Confusion Matrix
- Precision, Recall, F1

In credit risk, which is worse: approving a bad client (FN) or rejecting a good one (FP)? Explain.

---

# 5) Simple “Business Threshold”

Logistic regression outputs probabilities

### Compare 3 thresholds

Compute confusion matrix and metrics (f1, recall, precision) for for test set:

- Threshold = **0.65**
- Threshold = **0.50**
- Threshold = **0.35**

Explain:

- Which threshold you recommend and why
- How does lowering the threshold change the number of rejected clients?
- Which threshold reduces the number of bad clients approved?
- Which threshold increases the number of good clients rejected?
- If you were the Head of Risk, which threshold would you recommend and why?

---

# 6) Production-Oriented task: “Manual Review Queue”

In real banks, not all decisions are automatic.

### Task 6.1  Create 3 decision zones using probability

Using logistic regression probability p:

- **Auto-approve:** $p < 0.20$
- **Manual review:** $0.20 \leq p \leq 0.50$
- **Auto-reject:** $p > 0.50$

Deliverables:

- Count how many applications fall in each zone
- Business interpretation: how this reduces workload and risk

---

# 7) Business Demonstration

Create a final markdown section:

## Business Demo: Loan Approval Risk Model

Include:

1. **Executive summary**
- what dataset about
- what goal
- best model chosen
- key metric result
- recommended threshold
- how manual review queue works
2. **One table comparing models**
   Rows: models
   Columns: Accuracy, Recall (bad class), F1, comment (1 short sentence)
3. **One visual**
- Confusion matrix plot (matplotlib)
4. **Example: 3 applicants**
   Take 3 rows from test set and show:
- probability
- predicted decision zone (approve/review/reject)
- 1–2 sentence explanation for each (business style)