## **Exercice 1 : Définir le Problème et Collecter les Données – Prédiction des Défauts de Paiement sur les Prêts**

### ****Objectif du projet****

Accorder un prêt comporte toujours un risque : certains emprunteurs ne remboursent pas tout ou partie de ce qu’ils doivent. Pour une banque ou un organisme de crédit, ces défauts de paiement peuvent coûter très cher.

L’objectif de ce projet est de **mettre en place un système de prédiction automatique**, basé sur des données passées, afin d’estimer à l’avance si un emprunteur risque de ne pas rembourser son prêt. Cela permettra :

* de limiter les pertes financières,
* d’optimiser la prise de décision,
* et de proposer des conditions de prêt plus adaptées aux profils des clients.

### ****Quelles données faut-il pour prédire un défaut de paiement ?****

Pour bien anticiper les risques, il faut rassembler différents types de données sur les emprunteurs et leur comportement financier :

#### 👤 **1. Informations personnelles**

* Âge
* Sexe
* Situation familiale (marié, célibataire…)
* Niveau d’études
* Nombre d’enfants ou de personnes à charge
* Type de logement (locataire, propriétaire…)
* Durée de résidence actuelle
* Statut professionnel (salarié, indépendant, au chômage)

#### 💸 **2. Situation financière**

* Revenu mensuel ou annuel
* Montant des dettes en cours
* Ratio dette/revenu
* Solde des comptes bancaires
* Utilisation des cartes de crédit

#### 📈 **3. Historique de crédit**

* Score de crédit (type FICO ou autre)
* Retards ou incidents de paiement passés
* Ancienneté du crédit
* Nombre de crédits ouverts
* Nombre de demandes de crédit récentes

#### 🏦 **4. Détails du prêt demandé**

* Montant emprunté
* Durée du prêt
* Taux d’intérêt
* Objet du prêt (auto, logement, études…)
* Garantie ou bien mis en gage

#### **5. Historique de remboursement**

* Montants remboursés par rapport aux échéances prévues
* Retards de paiement
* Paiements partiels ou manqués
* Statut final du prêt (remboursé, en retard, défaut…)

#### **6. Données complémentaires (facultatives)**

* Paiements d’abonnements ou de factures (électricité, téléphonie…)
* Contexte économique global (inflation, chômage…)
* Données numériques (utilisation d’applications financières, si pertinent)

### ****Où récupérer ces données ?****

Voici les **principales sources** où l’on peut obtenir ce type d’informations :

#### **1. Données internes de l’organisme de prêt**

* Fiches de demande de prêt
* Données clients (revenus, historique bancaire…)
* Historique de remboursement

#### **🏦 2. Agences de crédit**

* Exemple : Equifax, Experian, TransUnion (ou CIBL, selon les pays)
* Fournissent des scores de crédit et des historiques de paiement

#### 🔄 **3. Partenaires externes**

* Plateformes de gestion de prêts
* API bancaires (open banking, avec autorisation du client)

#### **4. Bases publiques**

* INSEE, Banque Mondiale, ou autres instituts de statistiques pour les données économiques

#### **5. Autres fournisseurs de données (optionnels)**

* Fournisseurs d’électricité, téléphonie ou internet
* Données issues d’applications ou réseaux (dans un cadre légal et éthique)

## **Exercice 2 : Sélection des variables (features) et choix du modèle**

### ****Sélection des variables pertinentes****

Voici les variables (features) les plus susceptibles d’avoir un impact direct sur la probabilité qu’un emprunteur fasse défaut :

| **Catégorie** | **Variable** | **Pourquoi c’est important ?** |
| --- | --- | --- |
| Données personnelles | Âge | Les jeunes emprunteurs ont parfois moins de stabilité financière. |
|  | Situation professionnelle | Un emploi stable réduit le risque de défaut. |
|  | Revenu mensuel | Plus le revenu est élevé, plus la capacité de remboursement est forte. |
|  | Nombre de personnes à charge | Plus il y a de charges, plus le budget est serré. |
| Historique de crédit | Score de crédit | Mesure globale du comportement financier passé. |
|  | Nombre de crédits en cours | Trop de crédits = surcharge potentielle. |
|  | Retards de paiement passés | Excellent indicateur de comportement à risque. |
| Détails du prêt | Montant du prêt | Des montants élevés peuvent augmenter le risque. |
|  | Durée du prêt | Longue durée = risque plus élevé dans le temps. |
|  | Taux d’intérêt | Un taux élevé peut rendre le remboursement plus difficile. |
| Comportement de paiement | Retards récents | Indice direct de tension financière actuelle. |
|  | Paiements partiels ou manqués | Prédicteur immédiat du défaut. |

### ****2. Justification du choix des variables****

Les variables sont choisies pour ces 3 raisons principales :

* **Corrélation historique avec le défaut** : des études montrent que des antécédents de paiement, un score de crédit bas, ou un revenu faible sont parmi les meilleurs prédicteurs de défaut.
* **Pertinence métier** : un analyste crédit ou un agent de risque s’appuie déjà naturellement sur ces critères pour évaluer un dossier.
* **Disponibilité dans les données** : ces variables sont généralement disponibles dans les bases internes ou auprès des agences de crédit.

## **Exercice 3 : Entraîner, évaluer et optimiser le modèle de prédiction de défaut de prêt**

### ****Quel(s) modèle(s) choisir pour ce type de problème ?****

Comme il s'agit d’un **problème de classification binaire** (Défaut = Oui / Non), plusieurs modèles sont adaptés. Voici les plus pertinents :

#### **Modèles simples et interprétables**

1. **Régression logistique**
   * Facile à expliquer (bons pour les institutions financières)
   * Permet de comprendre l’impact de chaque variable
2. **Arbre de décision**
   * Interprétable visuellement
   * Pratique pour des règles simples

#### **Modèles plus puissants (mais moins transparents)**

1. **Random Forest**
   * Combine plusieurs arbres pour améliorer la précision
   * Réduit le surapprentissage
2. **XGBoost / LightGBM**
   * Très performants pour les problèmes de scoring
   * Gèrent bien les données déséquilibrées

En pratique : **Random Forest ou XGBoost** sont souvent les meilleurs compromis entre performance et robustesse pour ce cas d’usage.

### ****Étapes pour entraîner et évaluer le modèle****

#### 1**. Préparation des données**

* Nettoyage : suppression ou imputation des valeurs manquantes
* Encodage des variables catégorielles (one-hot, label encoding…)
* Normalisation des données numériques (si nécessaire)
* Séparation : **train/test split** (ex. 80/20)

#### 2. **Entraînement du modèle**

* Choisir un algorithme (ex. XGBoost)
* Lancer l'entraînement sur l’ensemble d'apprentissage
* (Optionnel) Réglage des hyperparamètres avec **Grid Search** ou **Random Search**

#### 3. **Évaluation des performances**

Puisqu’il s’agit d’un problème de **risque financier**, il est **crucial de bien détecter les défauts**, même s’ils sont rares.

## **Exercice 4 : Adapter le type d’apprentissage machine au problème**

#### **1. Prédire les prix des actions (Predicting Stock Prices)**

**Type de Machine Learning :**  
**Apprentissage supervisé (Supervised Learning)**  
**Type de tâche :**  
 **Régression**

**Explication :**   
Ce problème consiste à **prédire une valeur continue** (le prix futur d’une action), à partir de données historiques (prix passés, volume, indicateurs techniques, etc.).  
L'apprentissage supervisé est adapté car :

* On a des **données étiquetées** (prix connus à des dates précises),
* Le modèle apprend à prédire une **valeur numérique** à partir des tendances passées.

**Exemples de modèles utilisés :**

* Régression linéaire
* LSTM (réseaux de neurones récurrents pour séries temporelles)
* XGBoost

#### **📚 2. Organiser une bibliothèque de livres par genres ou similarités (Organizing a Library of Books)**

**Type de Machine Learning :**  
 **Apprentissage non supervisé (Unsupervised Learning)**  
**Type de tâche :**  
 **Clustering (regroupement)**

**Explication :**  
Dans ce cas, on veut **regrouper automatiquement des livres** selon leurs caractéristiques (extraits de texte, thème, mots-clés), **sans que les catégories soient connues à l’avance**.  
L’apprentissage non supervisé est idéal quand :

* Il n’y a **pas d’étiquette préexistante** (genre, catégorie),
* On veut découvrir des **structures cachées** dans les données.

**Exemples de techniques utilisées :**

* K-means
* DBSCAN
* Algorithmes de réduction de dimension (PCA, t-SNE)

***3. Programmer un robot pour naviguer dans un labyrinthe (Robot Maze Navigation)***

**Type de Machine Learning :**  
➡️ **Apprentissage par renforcement (Reinforcement Learning)**  
**Type de tâche :**  
➡️ **Agent autonome apprenant par essais/erreurs**

**Explication :**  
Ici, le robot doit **apprendre à prendre des décisions séquentielles** (aller à gauche, droite, reculer…) en interagissant avec un environnement.  
L'apprentissage par renforcement est utilisé quand :

* L'agent agit **en temps réel**,
* Il reçoit des **récompenses/punitions** (ex : +1 pour trouver la sortie, -1 pour heurter un mur),
* Le but est d’**optimiser une stratégie** (trouver le chemin le plus court).

**Exemples d’algorithmes utilisés :**

* Q-learning
* Deep Q Networks (DQN)
* Policy Gradient

## **Exercice 5 : Stratégie d’évaluation pour différents modèles d’apprentissage automatique**

#### **1Modèle supervisé (classification)**

**Exemple :** Prédire si un client va rembourser un prêt (**régression logistique**)

#### **L’Objectif :**

Évaluer dans quelle mesure le modèle **classifie correctement** les exemples en "défaut" ou "non défaut".**Stratégie d’évaluation :**

* **Méthodes :**
  + **Train/Test Split** (ex : 80/20) ou **K-fold Cross-Validation** (souvent k = 5 ou 10) pour évaluer la robustesse sur différents sous-ensembles.