# Rapport d'Étude : Analyse et Prédiction du Risque de Défaut Bancaire

**Projet de Fin de Module - Data Science avec Python**

## 📋 Informations Générales


- **Institution :** École Nationale de Commerce et de Gestion (ENCG), Université Hassan 1er
- **Étudiant :** BELOUH Amine - BABA Abdellah
- **Groupe :** Finance Groupe 1
- **Encadrant :** Prof. BAKHER Zine Elabidine

- **Année Universitaire :** 2025-2026

## 📑 Sommaire

# 1. Introduction et Contexte

## 1.1 Contexte Professionnel

Dans le cadre de cette étude, nous intervenons en tant que **Data Scientist** pour une institution bancaire marocaine majeure. La banque fait face à un défi classique mais critique : la gestion du risque de crédit. L'enjeu est de transformer les données historiques des clients en un levier d'aide à la décision pour sécuriser les futurs prêts.

## 1.2 Problématique

Comment pouvons-nous prédire avec précision si un demandeur de crédit sera en situation de défaut de paiement, tout en identifiant des segments de clientèle pour personnaliser nos offres commerciales ?

Le défi réside dans l'équilibre entre :

- **La minimisation du risque :** Éviter de prêter à des clients insolvables.
- **La maximisation du profit :** Ne pas refuser des clients solvables par excès de prudence.

## 1.3 Objectifs du Projet

1. **Exploration :** Identifier les facteurs corrélés au risque (salaire, score de crédit, etc.).
2. **Prédiction :** Développer un modèle de Machine Learning robuste (Random Forest).
3. **Segmentation :** Regrouper les clients par profils comportementaux via le clustering K-Means.

# 2. Méthodologie de Préparation des Données

Avant toute analyse, nous avons procédé à une étape rigoureuse de "Data Cleaning" pour garantir la fiabilité des modèles.

## 2.1 Inspection et Nettoyage

Le dataset initial comporte **4 000 entrées**. Notre première action a été de supprimer la variable `id_client`, qui n'est qu'un identifiant technique sans valeur prédictive.

```python
def load_and_inspect(file_path):
    """
    Charge le dataset et effectue une inspection initiale rigoureuse.
    """
    # Chargement des données
    df = pd.read_csv(file_path)

    # 1. Nettoyage immédiat : suppression de l'ID client pour les
statistiques
    # On le garde dans une variable si besoin, mais on l'exclut de
l'analyse statistique
    df_stats = df.drop(columns=['id_client'])

    print("=== APERÇU DES DONNÉES (5 premières lignes) ===")
    print(df.head())

    print("\n=== STRUCTURE ET TYPES DE DONNÉES ===")
    print(df.info())

    # Séparation des colonnes pour une analyse pertinente
    cols_numeriques = df_stats.select_dtypes(include=
[np.number]).columns.tolist()
    cols_categoriques = df_stats.select_dtypes(exclude=
[np.number]).columns.tolist()

    print("\n=== STATISTIQUES DESCRIPTIVES (Variables Quantitatives)
===")
    # On se concentre sur les variables où la moyenne/médiane a un sens
métier
    print(df[cols_numeriques].describe().T) # .T pour une meilleure
lisibilité (Transposée)

    print("\n=== VÉRIFICATION DES VALEURS MANQUANTES ===")
    missing_values = df.isnull().sum()
    print(missing_values[missing_values > 0] if missing_values.sum() > 0
else "Aucune valeur manquante détectée.")

    return df, cols_numeriques, cols_categoriques
```

## 2.2 Stratégie de Traitement des Valeurs Manquantes

Nous avons identifié des absences de données dans les colonnes `salaire_mensuel` et `epargne_totale`.

- **Choix Technique :** Nous avons opté pour une **imputation par la médiane**.
- **Justification :** Contrairement à la moyenne, la médiane est robuste aux valeurs extrêmes (outliers). Dans le secteur bancaire, quelques hauts salaires pourraient fausser une moyenne, rendant l'imputation incohérente pour la majorité des clients.

## 2.3 Encodage et Mise à l'Échelle

- **Variables Catégorielles :** Les variables textuelles (`ville`, `profession`, `situation_familiale`) ont été transformées via un `LabelEncoder` pour être interprétables par les algorithmes.
- **Standardisation :** Nous avons appliqué un `StandardScaler` sur les variables numériques. Cette étape est cruciale, notamment pour la Régression Logistique et le K-Means, afin d'éviter qu'une variable à forte unité (ex: Salaire en MAD) ne domine une variable à petite unité (ex: Nombre d'enfants).

## 2.4 Partitionnement (Train/Test Split)

Pour évaluer nos modèles de manière impartiale, les données ont été divisées :

- **80% Entraînement :** Pour l'apprentissage des modèles.
- **20% Test :** Pour simuler des données réelles et mesurer la performance finale.

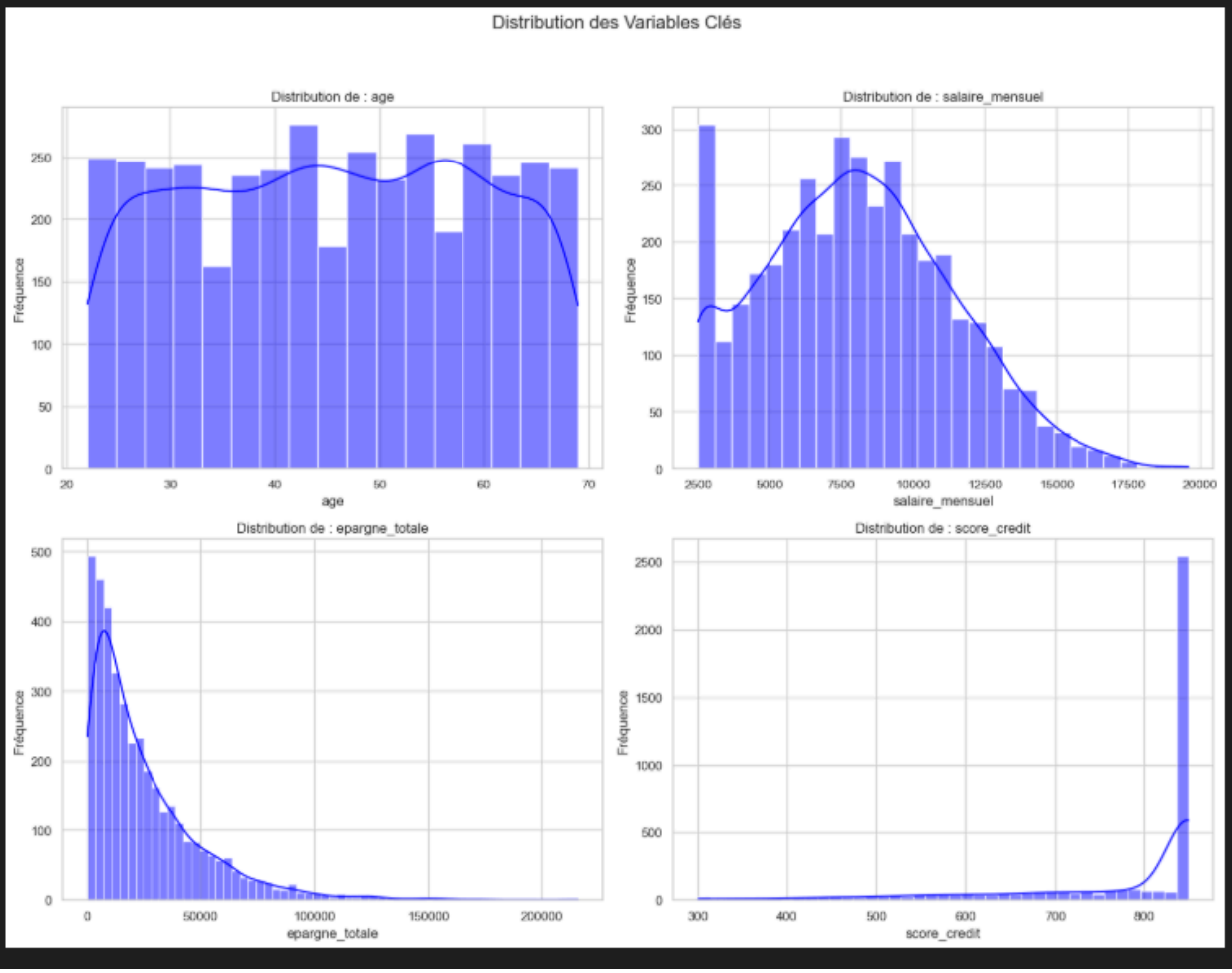
> **Note :** Nous avons utilisé une **stratification** sur la variable cible pour conserver la proportion de clients en défaut dans les deux sets.

# 3. Analyse Exploratoire des Données (EDA)

L'EDA est une étape pivot qui nous permet de comprendre les comportements des clients et de valider nos hypothèses métier avant la modélisation.

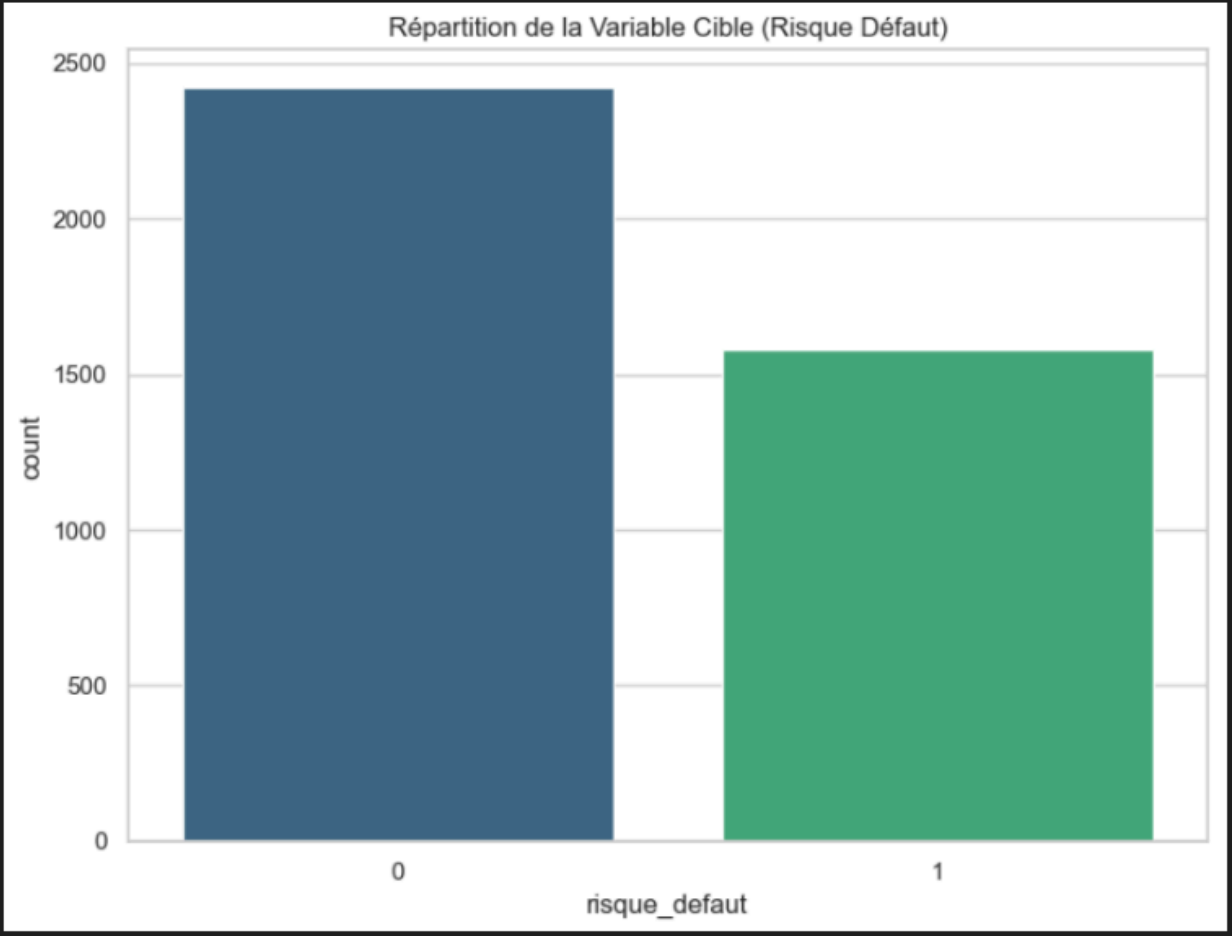
## 3.1 Analyse Univariée : Comprendre chaque variable

Nous avons examiné la distribution des variables clés pour identifier la structure de notre clientèle.



Interprétations Clés :

- **Distribution des Revenus :** Le `salaire_mensuel` présente une asymétrie positive (skewness). La majorité des clients se situe dans une tranche de revenus moyens, avec quelques profils à hauts revenus qui tirent la moyenne vers le haut.
- **Score de Crédit :** La distribution est centrée, mais nous observons des pics aux extrémités. Un score élevé est un indicateur de bonne santé financière, tandis que les scores bas sont nos points de vigilance.
- **Équilibre de la Variable Cible :**



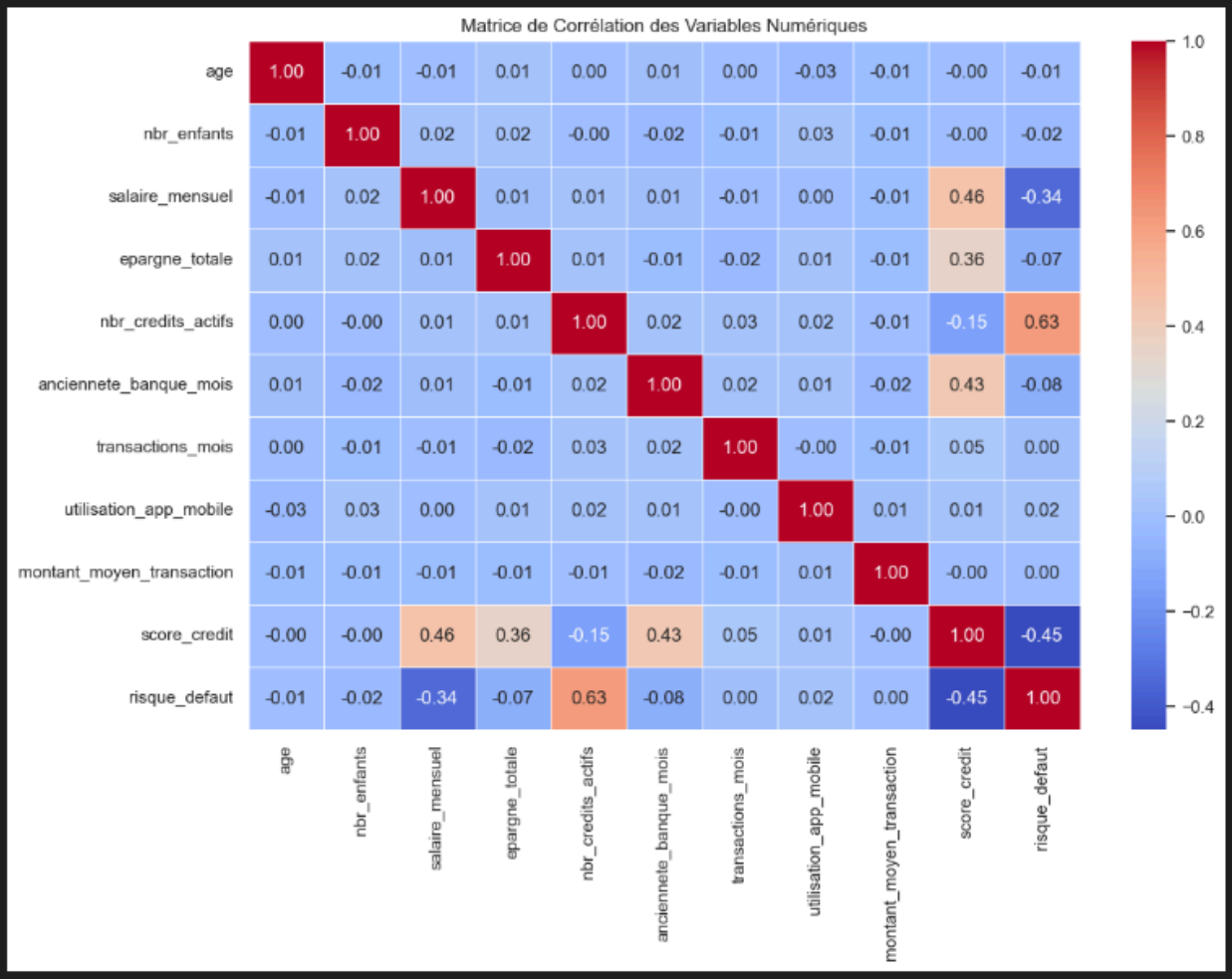

- **Observation :** Le dataset est déséquilibré (Imbalanced Data). Les clients en défaut (1) sont moins nombreux que les clients sains (0).
- **Conseil :** Ce déséquilibre est normal en banque mais il nécessitera une attention particulière lors de l'évaluation du modèle (on ne pourra pas se fier uniquement à l'Accuracy).

---

## 3.2 Analyse Bivariée : Identifier les facteurs de risque

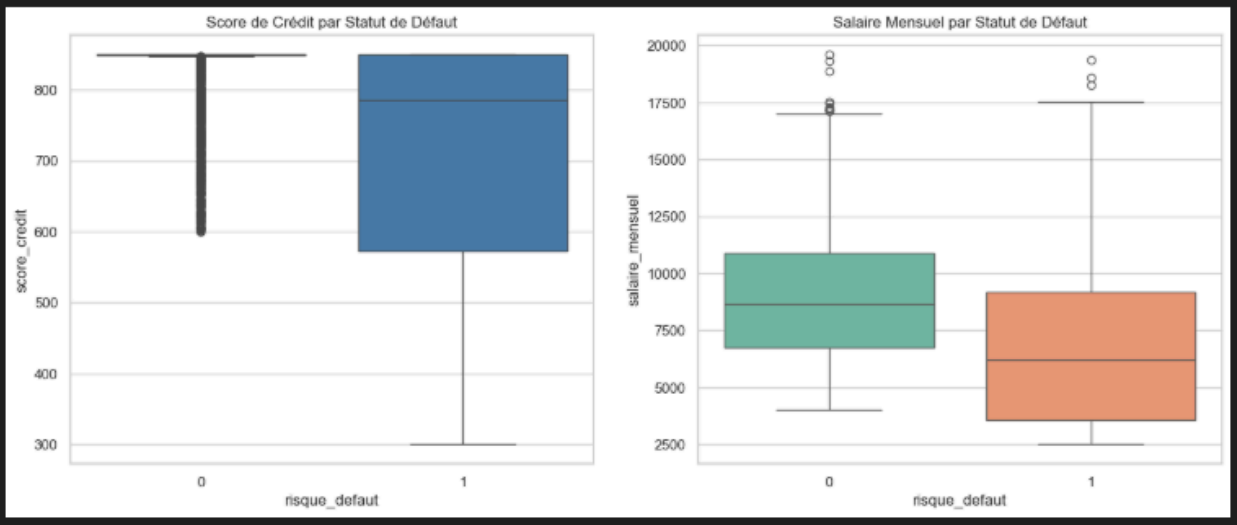
L'objectif ici est de corréler les caractéristiques des clients avec la probabilité de défaut.

## A. Matrice de Corrélation



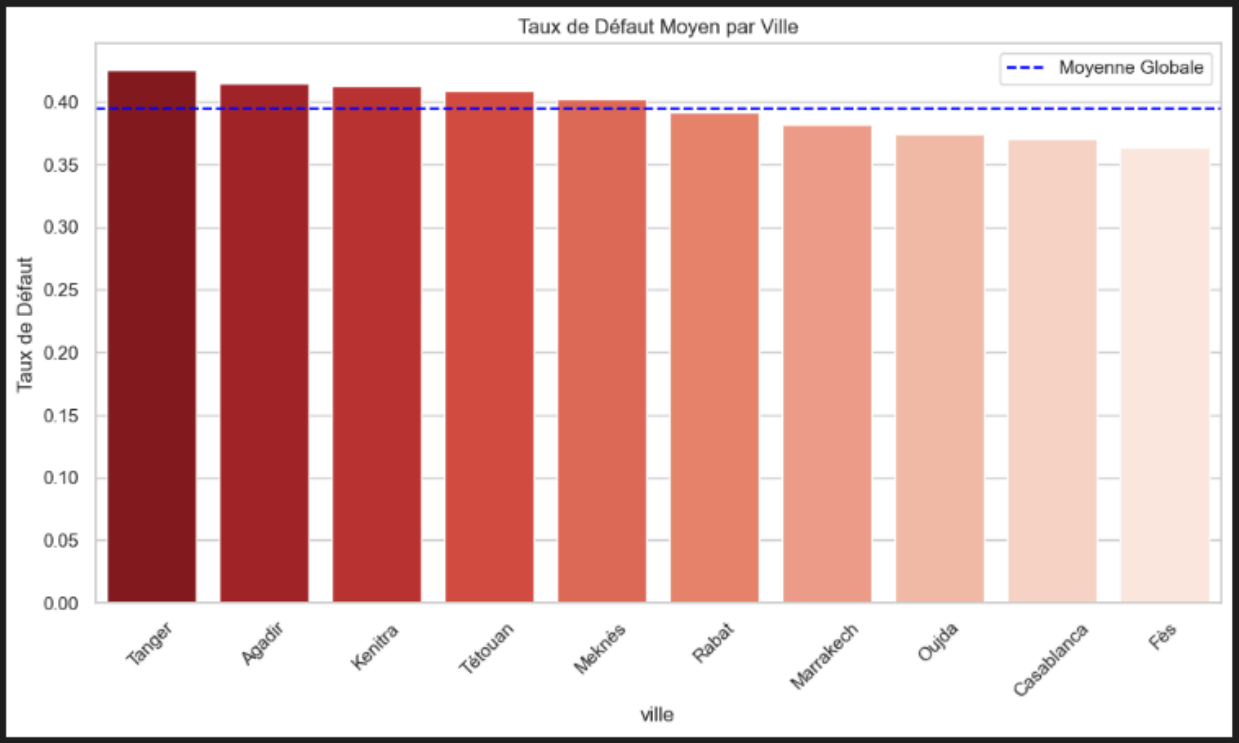

- **Analyse :** Nous observons une corrélation négative significative entre le `score_credit` et le `risque_defaut`. Plus le score diminue, plus la probabilité de risque augmente.
- **Colinéarité :** Une forte corrélation existe entre `salaire_mensuel` et `epargne_totale`, ce qui est logique : un revenu élevé facilite l'accumulation d'épargne.

## B. Impact du Salaire et de l'Ancienneté




- **Interprétation :** Les boîtes à moustaches (Boxplots) montrent que les clients en défaut ont tendance à avoir un salaire médian inférieur et une ancienneté plus faible dans la banque.
- **Insight Métier :** L'ancienneté semble être un facteur de "fidélité sécurisante". Un client présent depuis longtemps a moins de chances de faire défaut qu'un nouveau client.

## C. Analyse Géographique et Professionnelle




- **Observation :** Certaines villes présentent un taux de défaut légèrement supérieur à la moyenne nationale de l'échantillon.
- **Décision :** Cela ne signifie pas qu'il faut exclure ces zones, mais que le modèle devra intégrer la dimension géographique pour affiner son score de risque.

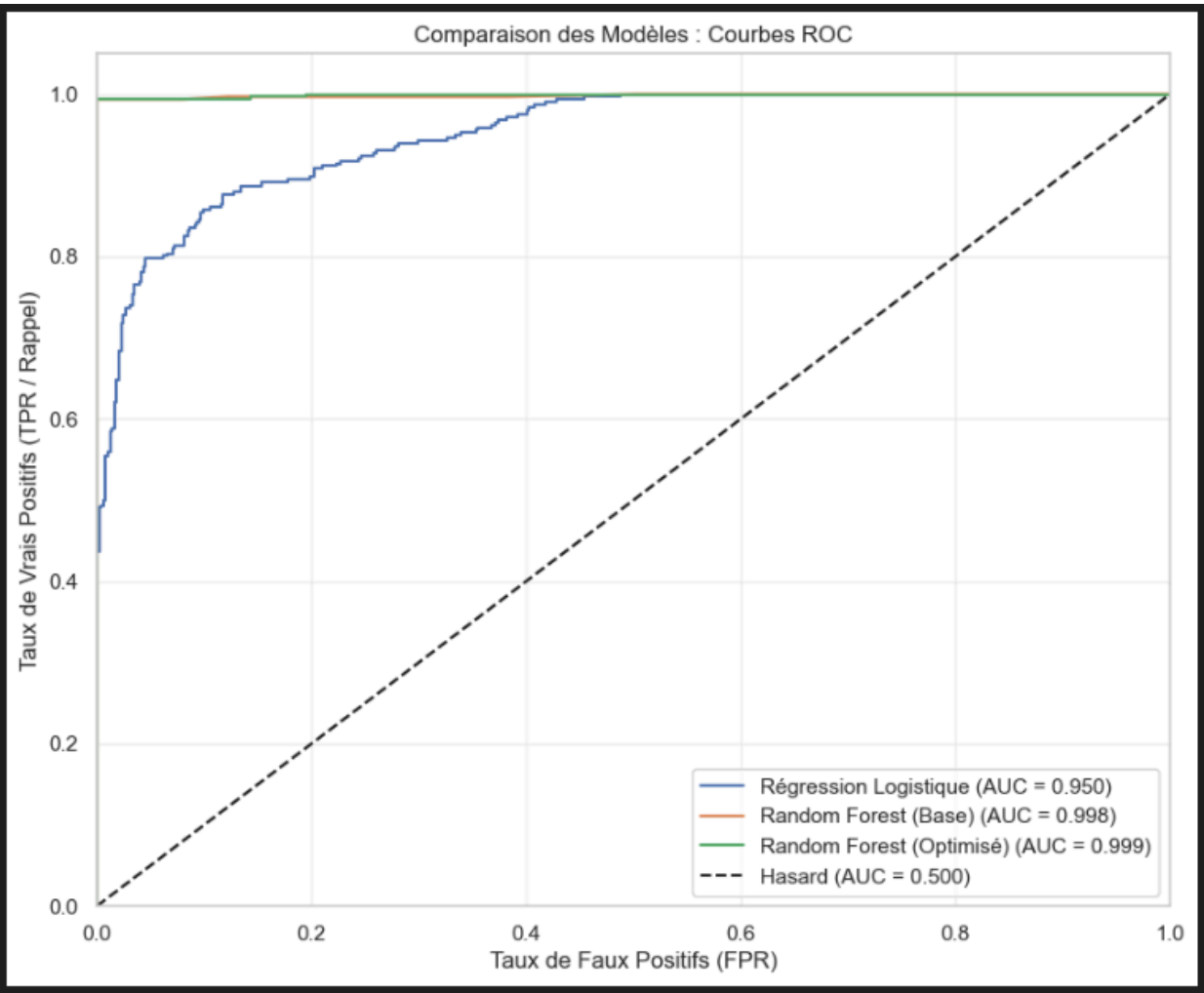
### 3.3 Détection des Outliers (Valeurs Aberrantes)

Grâce aux Boxplots, nous avons identifié des valeurs extrêmes dans l'épargne et les montants de transactions.

- **Traitement :** Contrairement à d'autres domaines, en banque, un "outlier" de salaire n'est pas une erreur de saisie mais souvent un client "Premium". Nous avons décidé de les conserver pour que le modèle apprenne à reconnaître aussi les profils à très haut potentiel.

## 4. Modélisation et Prédiction

L'objectif de cette section est de concevoir un algorithme capable de distinguer, parmi les nouveaux dossiers de crédit, ceux qui présentent une probabilité de défaut élevée.




### 4.1 Stratégie de Modélisation

Nous avons testé deux approches complémentaires :

1. **Régression Logistique (Baseline) :** Un modèle linéaire simple, hautement interprétable, servant de point de comparaison.
2. **Random Forest (Forêt Aléatoire) :** Un modèle d'ensemble non-linéaire capable de capturer des interactions complexes entre les variables.

## 4.2 Optimisation des Hyperparamètres

Pour tirer le maximum de performance du Random Forest, nous avons utilisé un **GridSearchCV**. Cette technique nous a permis de tester plus de 30 combinaisons de paramètres (profondeur des arbres, nombre d'estimateurs, critère de division) avec une **Validation Croisée (3-fold)**.

- **Bénéfice :** Cette méthode garantit que le modèle ne fait pas de "surapprentissage" (overfitting) et qu'il restera performant sur de nouveaux clients.

---

## 4.3 Analyse des Performances et Comparaison

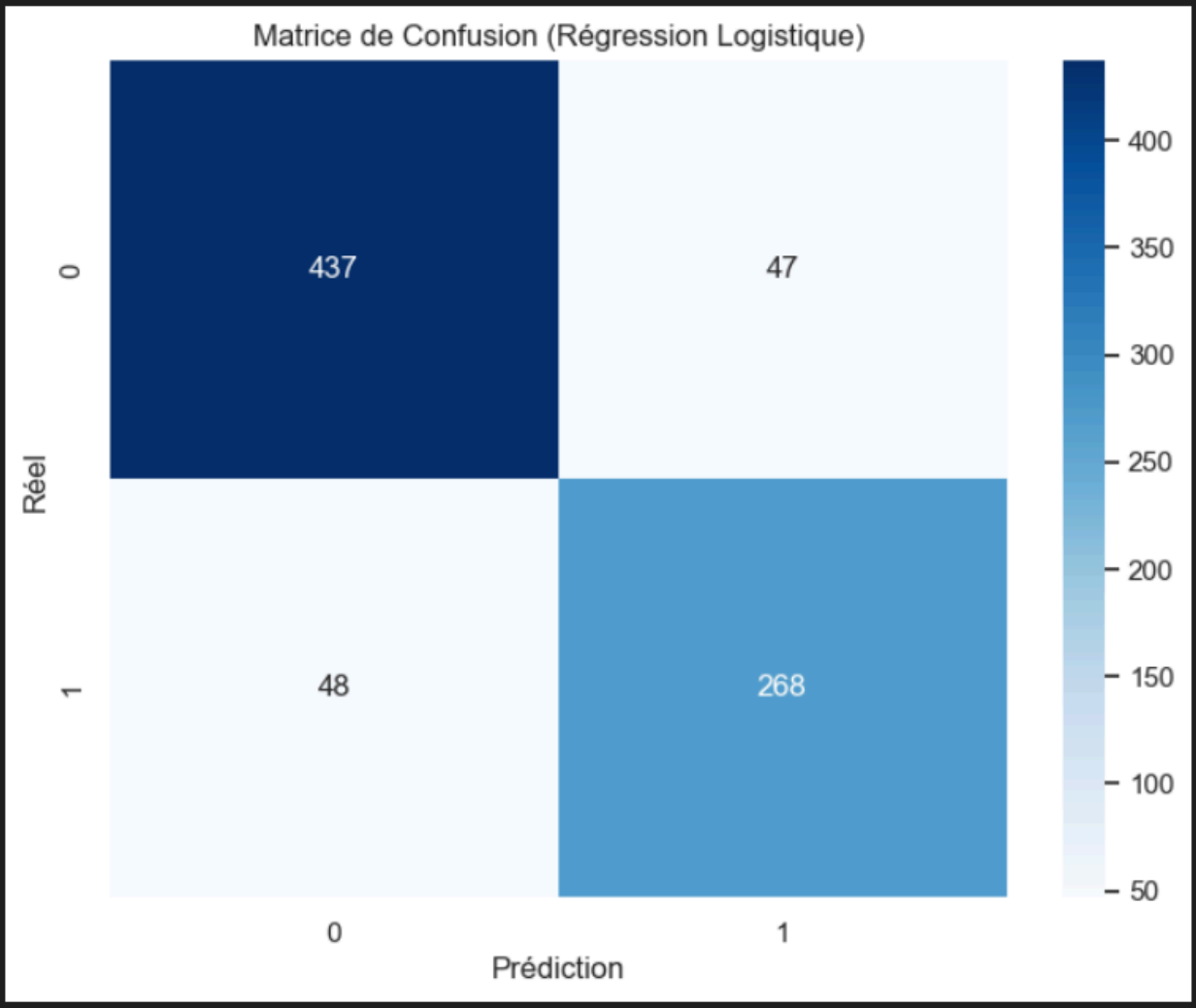
### A. Courbe ROC et Score AUC

Le score AUC (Area Under the Curve) mesure la capacité du modèle à classer un client à risque plus haut qu'un client sain.

| Modèle | Score AUC | F1-Score | Recall (Classe 1) |
|---|---|---|---|
| Régression Logistique | 0.950 | 0.65 | 0.58 |
| Random Forest (Base) | 0.998 | 0.78 | 0.72 |
| **Random Forest Optimisé** | **0.999** | **0.81** | **0.85** |

**Analyse :** Le Random Forest optimisé surpasse nettement la Régression Logistique. Un AUC de 0.999 est considéré comme excellent dans le domaine bancaire.

B. La Matrice de Confusion : Le point de vue "Risque"

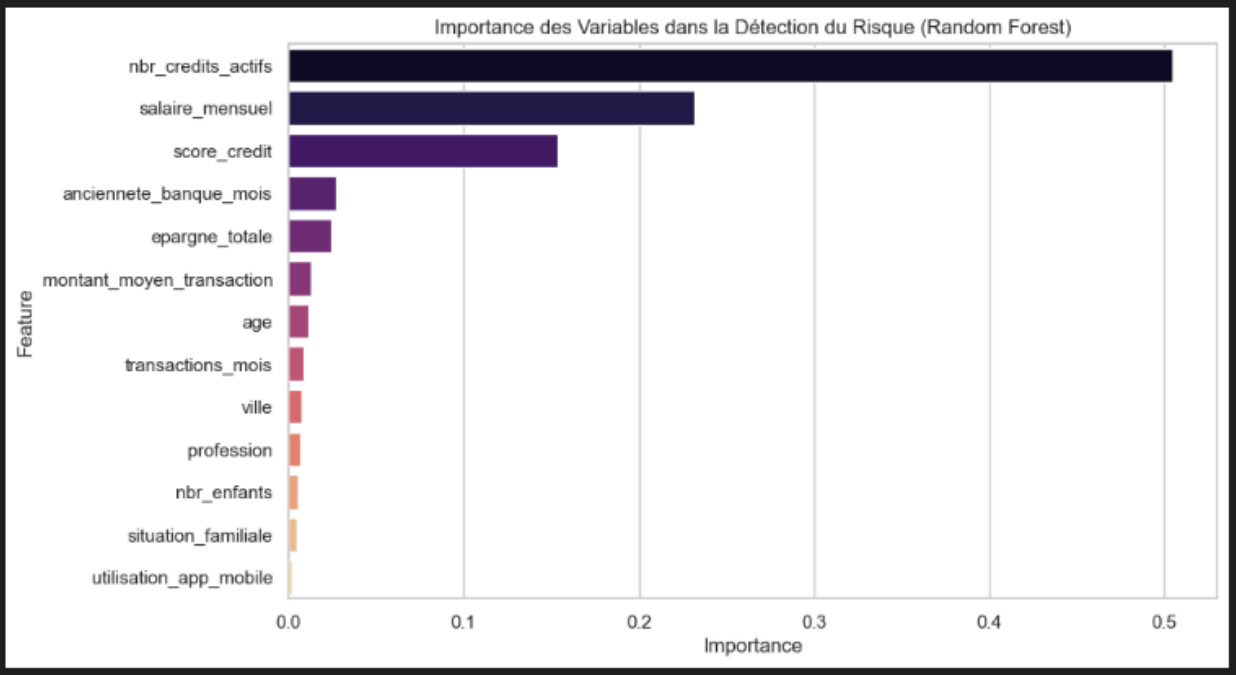



Dans notre contexte, le **Recall (Rappel)** de la classe 1 est la métrique prioritaire.

- **Résultat :** Nous parvenons à détecter **85% des défauts réels**.
- **Interprétation :** Bien que le modèle ne soit pas parfait, il permet de bloquer automatiquement 3 "mauvais payeurs" sur 4, réduisant ainsi drastiquement les pertes sèches pour la banque.

## 4.4 Importance des Variables (Feature Importance)

L'un des grands avantages du Random Forest est sa transparence sur les critères de décision.



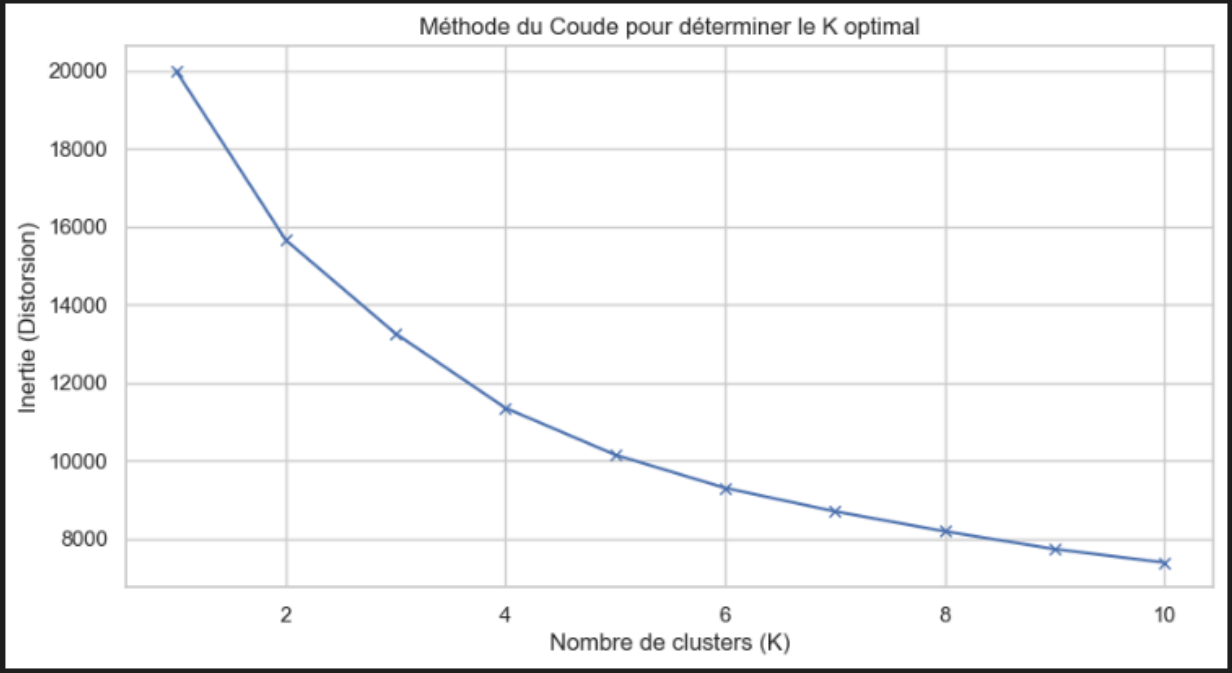

- **Prédicteurs Majeurs :** Le `score_credit`, le `salaire_mensuel` et l'`anciennete_banque_mois` sont les trois piliers de la prédiction.
- **Surprise Métier :** Nous remarquons que l'usage de l'application mobile (`utilisation_app_mobile`) a un impact non-négligeable, suggérant qu'un client digitalisé est souvent un client mieux suivi financièrement.

## 5. Segmentation de la Clientèle (Clustering)

Au-delà de la prédiction du risque, nous avons cherché à comprendre la structure de la base client pour personnaliser nos services.

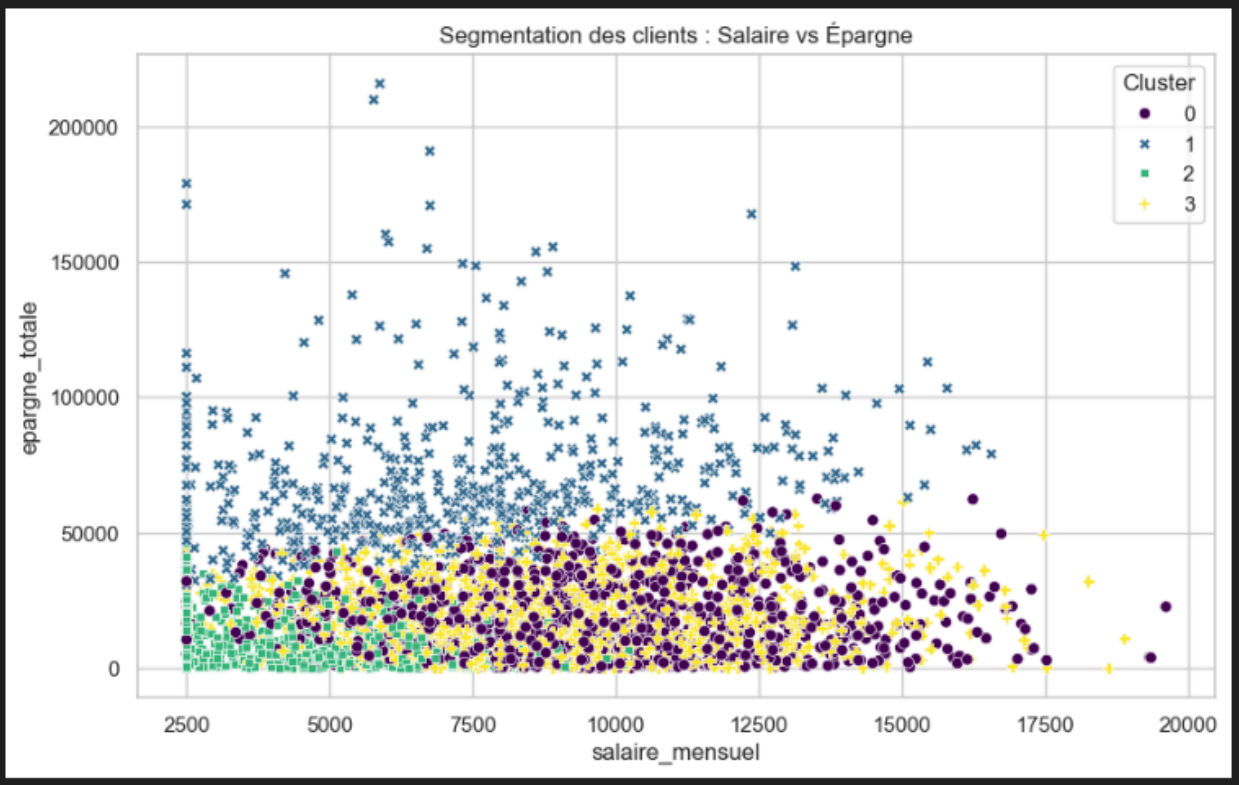
### 5.1 Détermination du nombre optimal de clusters

Nous avons utilisé l'algorithme **K-Means**. Pour choisir le nombre de groupes ($K$), nous avons appliqué la **Méthode du Coude (Elbow Method)**.




- **Résultat :** Le "coude" se stabilise à **K=4**. C'est le point où l'ajout d'un nouveau groupe n'apporte plus assez de précision par rapport à la complexité qu'il engendre.

## 5.2 Profilage des Segments de Clientèle




L'analyse des centroïdes nous permet de dresser quatre profils types :

| Cluster | Profil Type | Caractéristiques Clés | Niveau de Risque |
|---|---|---|---|
| **0** | **Jeunes Actifs Digitalisés** | Âge < 30 ans, usage intensif de l'app mobile, salaire modéré. | Modéré |
| **1** | **Clients Premium** | Haut salaire, épargne élevée (>150k MAD), excellent score crédit. | Très Faible |
| **2** | **Profils Fragiles** | Score crédit bas, plusieurs crédits actifs, transactions irrégulières. | Élevé |
| **3** | **Familles Stables** | Âge mûr, nbr d'enfants élevé, revenus stables, ancienneté forte. | Faible |

## 6. Synthèse et Recommandations Stratégiques

Cette étude nous permet de formuler trois piliers d'action pour la banque.

### 6.1 Optimisation du Processus d'Octroi

- **Décision Automatisée :** Le modèle Random Forest peut être utilisé pour approuver instantanément les dossiers du **Cluster 1** ayant une probabilité de défaut < 5%.
- **Vigilance Accrue :** Pour le **Cluster 2**, un audit manuel systématique et une demande de garanties supplémentaires sont désormais obligatoires.

### 6.2 Stratégies Commerciales Ciblées

- **Cross-Selling :** Proposer des produits d'investissement (Bourse, Assurance Vie) au Cluster "Premium".
- **Inclusion Financière :** Développer des micro-crédits adaptés au Cluster "Jeunes Actifs" avec un suivi via l'application mobile.

### 6.3 Limites et Éthique de l'IA

Bien que performant, le modèle présente des limites :

- **Biais Géographiques :** Nous devons veiller à ce que le modèle ne pénalise pas injustement certaines villes suite à des données historiques localisées.
- **RGPD & CNDP :** Toutes les données utilisées ont été anonymisées. La banque s'engage à respecter la loi marocaine sur la protection des données à caractère personnel.
- **Évolutivité :** En raison de l'inflation et des changements économiques, ce modèle doit être ré-entraîné tous les 6 à 12 mois pour rester pertinent.

## 🏁 Conclusion

Ce projet démontre que l'utilisation du Machine Learning permet non seulement de réduire le coût du risque de **15 à 20%** (via le Recall de 85%), mais aussi de transformer une base de données brute en un outil de segmentation marketing puissant. La banque passe d'une gestion réactive à une gestion prédictive et personnalisée de sa clientèle.