Rédigé par : Angèle Blandine Feussi Nguemkam -Willy Stanlin Taguedong - Thierry Pascal Zokou Tchokonthe - Sorel Aniel Fotsing Mba

RAPPORT FINAL – Analyse Exploratoire Multivariée (UA3)

1. Introduction

Ce projet vise à analyser les facteurs associés au défaut de paiement à partir du dataset UCI Credit Card Default, qui réunit les caractéristiques de 30 000 titulaires de cartes de crédit. À travers un ensemble de méthodes statistiques – ACP, AFC, ACM et K-Means – l'objectif est d'identifier les déterminants majeurs du risque client et de proposer une segmentation exploitable pour un futur modèle de scoring. Le travail inclut également un prétraitement rigoureux garantissant la cohérence des données.

2. Description du dataset

Le jeu de données comporte 24 variables, regroupées en trois familles :

- Socio démographie : sexe, niveau d'éducation, situation matrimoniale, âge.
- Comportements de paiement : retards mensuels (pay_0 à pay_6), montants facturés (bill_amt1–6), montants payés (pay_amt1–6).
- Variable cible : default_payment_next_month (default = 1).

3. Prétraitement et validation

Une phase de nettoyage a permis d'assurer la qualité des données. Les variables catégorielles ont été vérifiées afin de respecter les modalités autorisées. Les montants facturés et payés ont été inspectés et corrigés pour éliminer les valeurs aberrantes ou infinies. Les données numériques ont été standardisées afin de permettre un ACP et un clustering fiables. Ces étapes garantissent une base de données cohérente et conforme aux règles métier.

4. Analyse univariée et bivariée

Les résultats montrent que les caractéristiques sociodémographiques influencent très peu le défaut. En revanche, les comportements de paiement constituent un révélateur essentiel du risque. Plus les retards sont importants, plus le taux de défaut augmente, notamment au-delà de deux mois d'arriérés. L'historique des paiements apparaît donc comme la principale source de variabilité, observation qui sera confirmée par les analyses multivariées.

5. Analyse Factorielle des Correspondances (AFC)

L'AFC appliquée aux variables catégorielles révèle que le sexe, l'éducation et la situation matrimoniale ne différencient pas les clients selon leur risque. À l'inverse, les retards (pay_0 à pay_6) créent une séparation nette entre les clients en défaut et les bons payeurs. Les modalités associées à plusieurs mois de retard se situent clairement à proximité de la modalité « défaut ». L'AFC confirme ainsi que le risque client est structuré par les comportements de remboursement plutôt que par les caractéristiques personnelles.

6. Analyse en Composantes Principales (ACP)

L'ACP indique que deux axes dominent la structure des données. Le premier (≈ 40 % de variance) agrège les montants facturés et payés, tandis que le second (≈ 12 %) isole les comportements atypiques. La variable cible se projette près des vecteurs pay_0 à pay_6, illustrant une forte corrélation entre le défaut et les retards. Les montants financiers contribuent à la structure générale des données, mais expliquent peu le risque. L'ACP valide donc une fois de plus la centralité du comportement de paiement dans l'analyse du défaut.

7. Analyse des Correspondances Multiples (ACM)

L'ACM enrichit l'interprétation en montrant une dispersion marquée des modalités de retard, qui se regroupent selon leur intensité. Les clients payant à temps s'opposent nettement à ceux accumulant plusieurs mois de retard. L'ACM confirme que le risque est d'abord un comportement avant d'être un profil démographique, ce qui renforce les conclusions tirées de l'AFC et de l'ACP.

8. Clustering – KMeans

La classification K-Means avec k = 4 permet d'obtenir une segmentation plus nuancée des comportements de paiement. Un premier cluster regroupe les bons payeurs, sans retards et présentant un risque de défaut très faible. Un second cluster rassemble des clients relativement stables, avec quelques retards occasionnels mais des montants maîtrisés, indiquant un risque faible à modéré. Le troisième cluster identifie des clients plus fragiles, caractérisés par des retards plus fréquents et des montants facturés plus élevés, révélant une situation financière tendue. Enfin, un quatrième cluster correspond aux profils les plus vulnérables, cumulant retards prolongés, montants dus importants et taux de défaut très élevé. Cette segmentation en quatre groupes offre une lecture plus détaillée du portefeuille client et constitue un outil opérationnel pour adapter les stratégies de gestion du risque et affiner un modèle de scoring.

9. Conclusion

Les quatre méthodes convergent vers une même conclusion : l'historique des retards est le facteur déterminant du défaut de paiement, tandis que les variables financières structurent secondairement les comportements et que les variables sociodémographiques jouent un rôle marginal. Le projet permet ainsi de dégager une compréhension approfondie du risque client et constitue une base solide pour la mise en place d'un modèle prédictif ou d'une stratégie de segmentation automatisée.