IF29-Traitement de données (Data Analytics)

Comparaison de deux méthodes de classification de profils de X (ex. Twitter)

Membres du groupe

- LATH Victor
- MBANGUE Patrice
- MOHAMMAD Ahamad
- NEIL-JOVY Minko
- NGUEMO KAMWOUA Dora
- TAKAM TALLA Vigny

Superviseurs: BIRREGAH Babiga, GUEPIE Blaise

Printemps 2025

Sommaire

Ingestion des données

Feature engineering

Analyse des données

Méthode de labélisation des données

Modélisation

Conclusion: Comparaison des résultats

INTRODUCTION

- Surveillance des réseaux sociaux → présence de bots/spammeurs

- Objectif : Détecter les profils atypiques via deux approches :
  - Apprentissage supervisé ()
  - Apprentissage non-supervisé ()

- Comparaison des performances et pertinence

INGESTION DES DONNEES

Ingestion de données

Pourquoi?

- 4.6 Millions de documents de tweets
- Mongodb système de gestion des bases de données orientés documents
- Création des pipelines de données

Hypothèse à prendre en compte lors de la création des pipelines des données :

- Trie du dataset suivant la date de création
- Extraction des informations les plus récentes des users(user_name, user_location, user_description, user_followers_count, user_statuses_count, user_created_at et user_verified …)
- Agrégation des données suivant user_id
- Récupération des informations notamment la date du tweet le plus récent(qui servira après pour le calcul de l'age du compte)

- Identifiants et Informations de Profil
- Statistiques du Compte Utilisateur
- Statistiques sur les Tweets
- Indicateurs Moyens et Ratios
- Variables Dérivées (Booléennes)

FEATURE ENGINEERING

Calcul des Nouvelles Caractéristiques

- Degré d'agressivité
- Ratio nombre d'abonnés/nombre de comptes suivis
- Ratio nombre de comptes suivis/nombre d'abonnés
- Score de visibilité

ANALYSE DES DONNEES

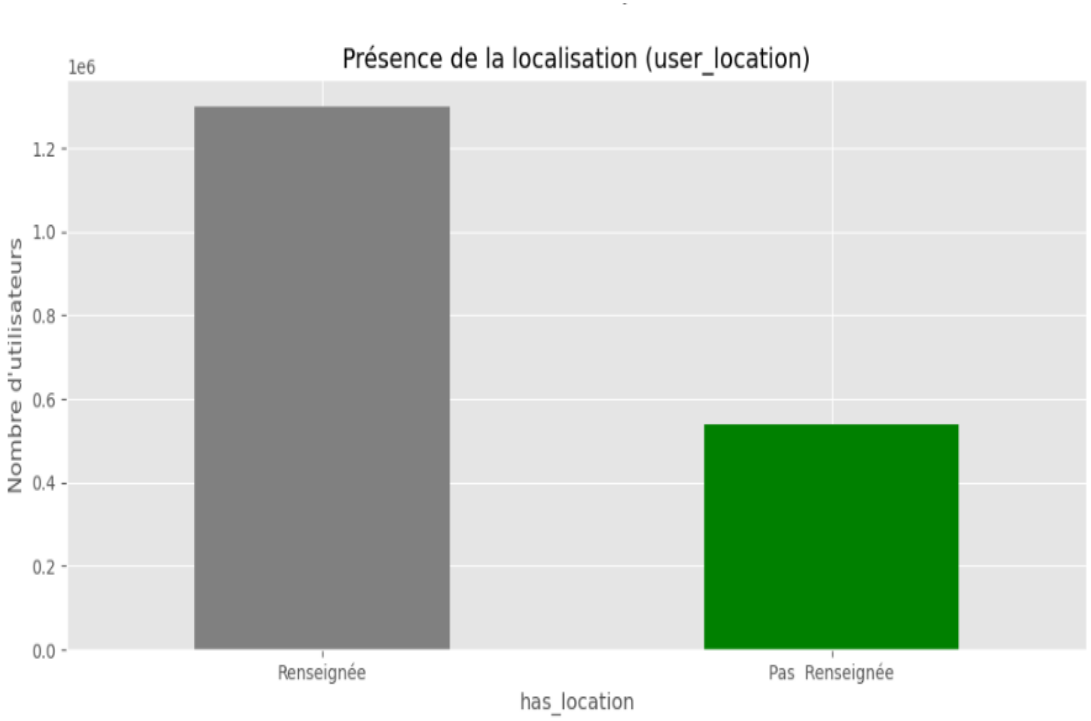
Analyse des données

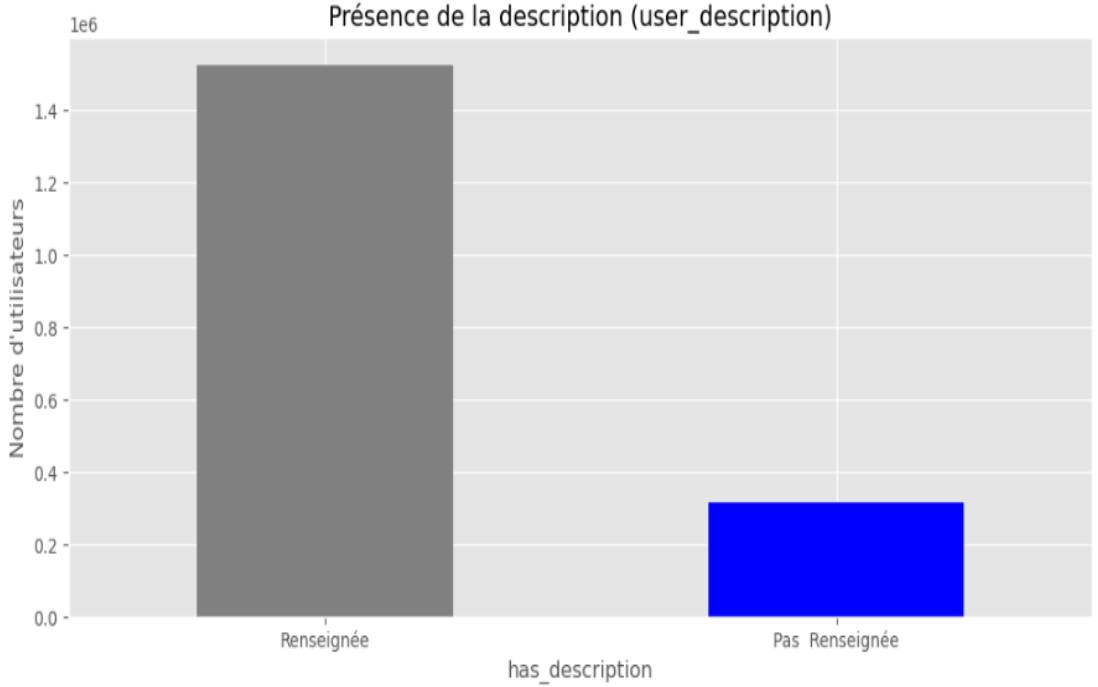
Analyse descriptive des valeurs manquantes

_id	0.000000
user_name	0.000054
user_location	29.288739
user_description	17.127960
user_followers_count	0.000000
user_friends_count	0.000000
user_statuses_count	0.000000
user_created_at	0.000000
user_verified	0.000000
user_favourites_count	0.000000
user_default_profile	0.000000
user_default_profile_image	0.000000
user_listed_count	0.000000
user_lang	0.000000
user_source	0.000000
tweets	0.000000
total_tweets	0.000000
total_retweets	0.000000
total_favorites	0.000000
total_retweeted_tweets	0.000000

Analyse des données

Analyse descriptive des valeurs manquantes





Approches exploitées

- Suppression des données
  - Conséquence: Perte d'information considérables

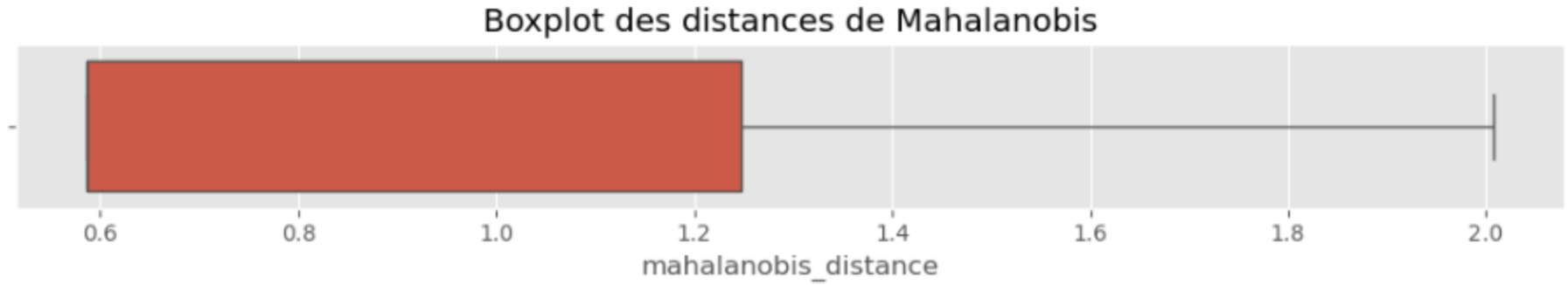
- Amputation par le mode
  - Calcul du mode pour les deux variables

```
Mode de user_location : 'Lagos, Nigeria' (0.88%) des valeurs non nulles
Mode de user_description : '.' (0.07%) des valeurs non nulles
```

  - Conséquence: Mode très faiblement représenté, et risque des sur-représentativité si on procède à l'amputation par le mode

- Solution
  Amputation par la nouvelle modalité "Non renseignée"

- **Vérification et traitement de valeurs extrèmes en utilisant la distance de Mahalanobis**



Interprétation: L'ensemble des observations semble donc cohérent et homogène par rapport à la structure globale du dataset. Aucune suppression ou correction n'est nécessaire à ce stade.

METHODE DE LABELISATION DES DONNEES ET ANALYSE

Méthode de labélisation des données et analyse

Définition du profil suspect à partir des caractéristiques utilisateurs 1/2

- **Profil non personnalisé** : photo ou thème par défaut (`user_default_profile_image` ou `user_default_profile = True`) → typique des faux comptes.

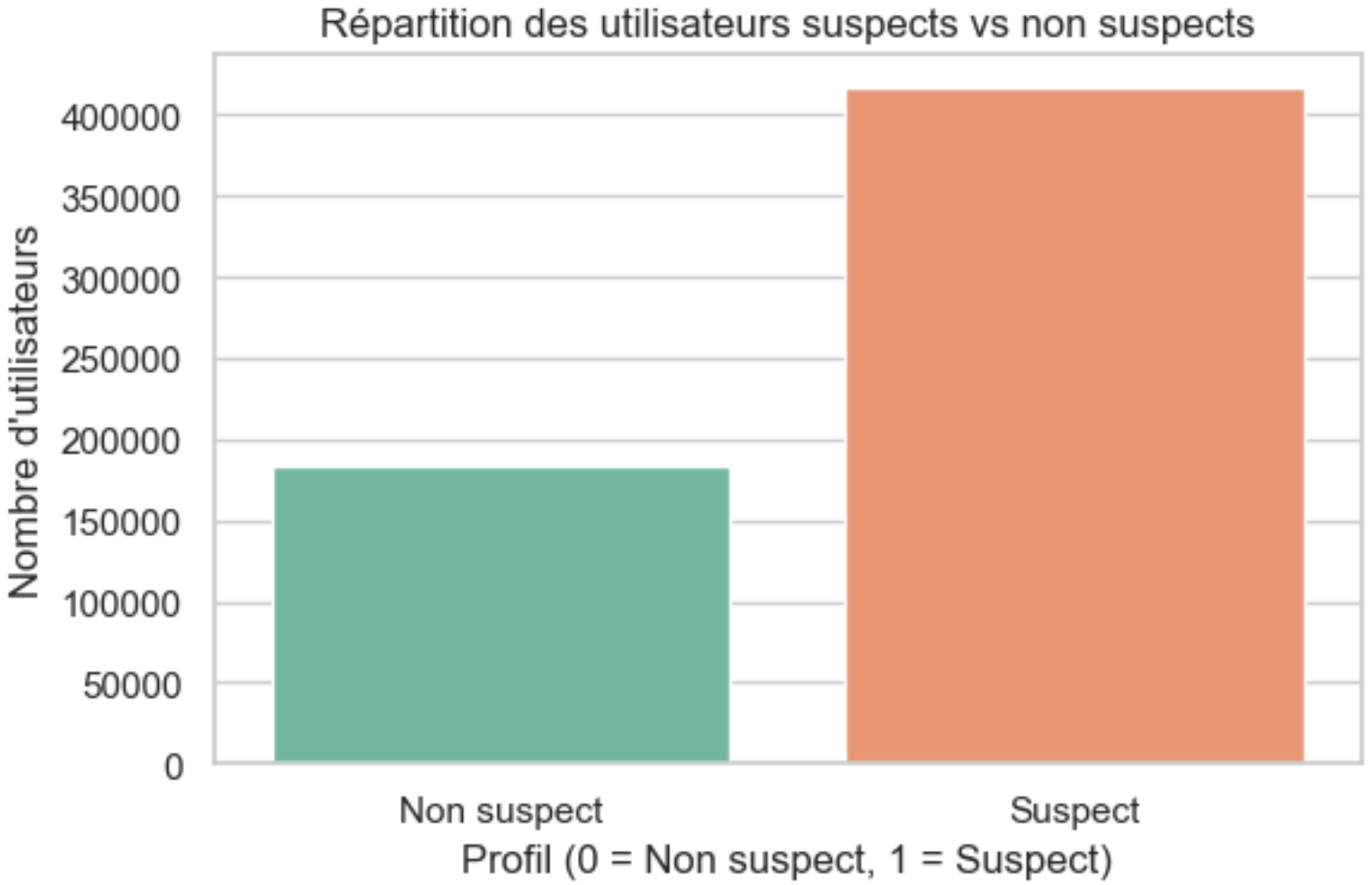
- **Activité anormalement élevée** : activité très élevée par rapport à l'âge du compte (`degree_agressiv > 95e percentile`).

- **Ratio d'abonnements déséquilibré** :suit beaucoup mais peu suivi (`friends_follow_ratio > 10` ou `ratio_followers_friends < 0.1`).

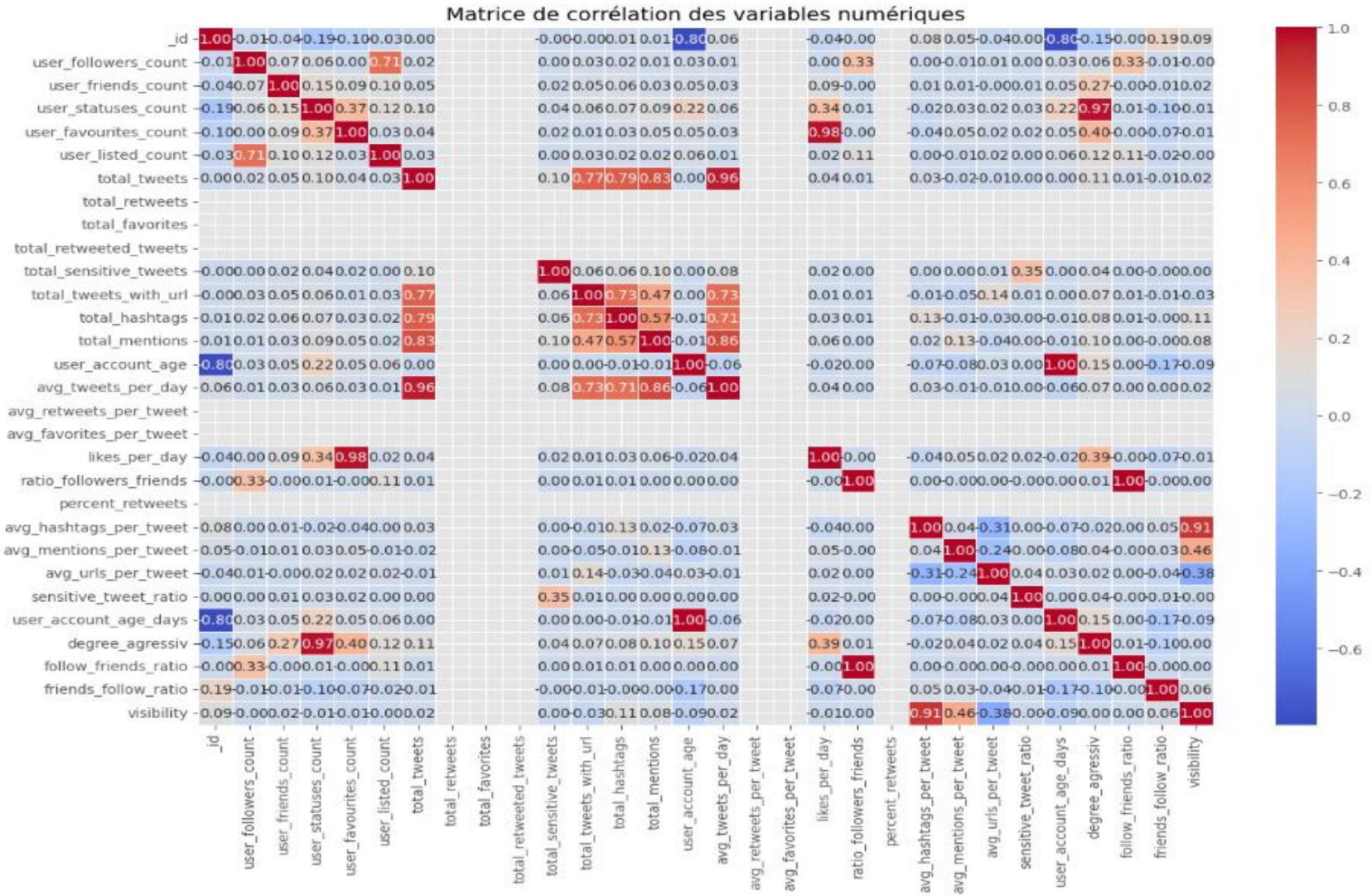
- **Contenus sensibles** : grande part de tweets marqués sensibles (`sensitive_tweet_ratio > 0.5`).
- **Spamming** : trop de liens dans les tweets (`avg_urls_per_tweet > 1`).
- **Profil incomplet** : absence de bio ou de localisation (`has_description` ou `has_location = "Non renseignée"`).
- **Compte non vérifié mais hyperactif** : compte non vérifié avec activité journalière excessive (`verified = False` et `avg_tweets_per_day > 95e percentile`)

Méthode de labelisation des données et analyse

Analyse du dataset labelisé



Matrice de corrélation des variables



MODELISATION

- **Apprentissage supervisé: SVM (Support Vector Machine)**

  - Bonne généralisation
  - Efficace après réduction de dimension
  - Moins de réglages

  Limite : il peut être lent sur de très grands volumes

- **Apprentissage non supervisé : K-Means**

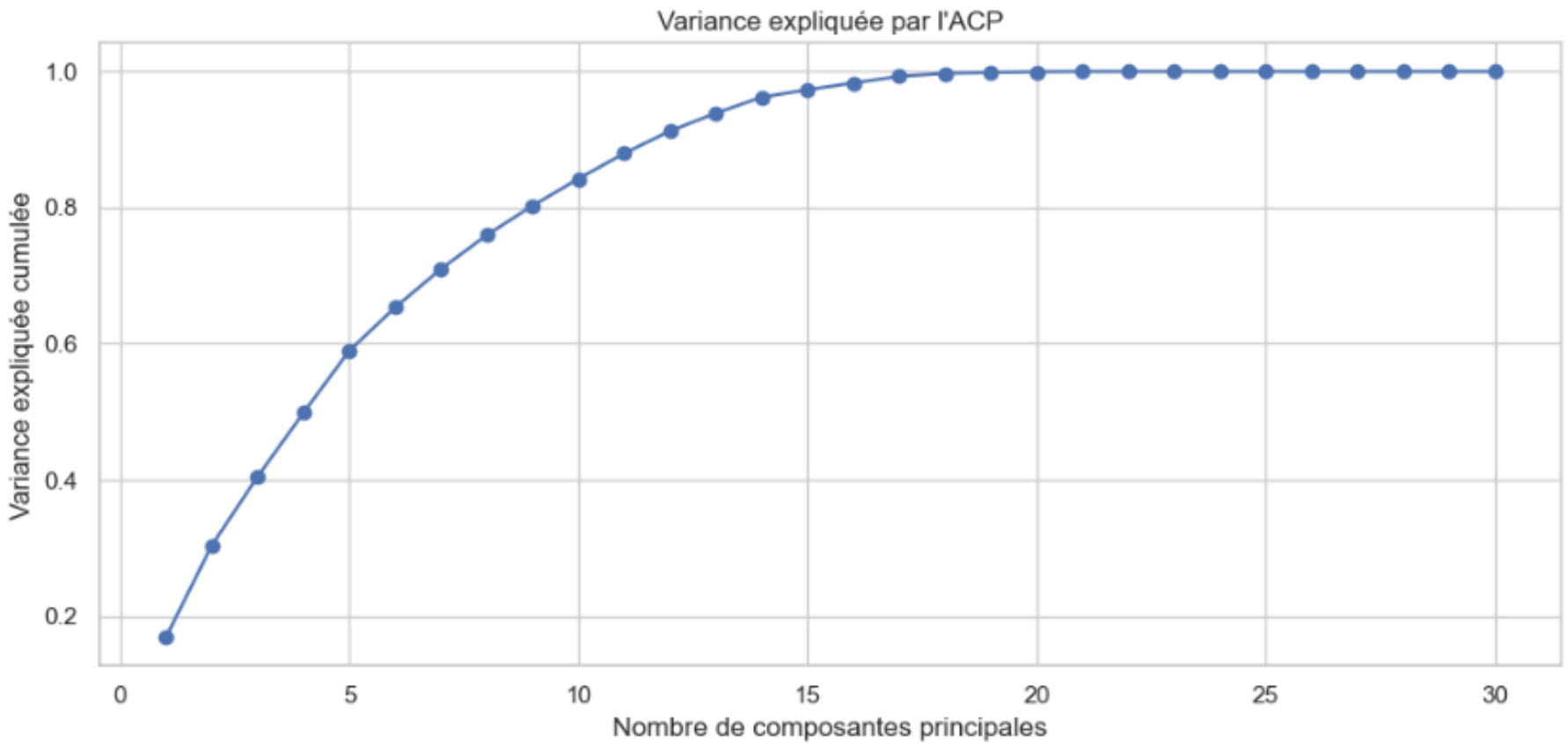
  - Rapide et efficace
  - Identification de profils
  - Adapté à nos données
  - Interprétable

Première approche:

Apprentissage supervisé: SVM (Support Vector Machine)

ACP pour la reduction de la dimension

Variance expliquée par l'ACP



Nombre de composantes principales à conserver pour expliquer au moins 80% de la variance : 9

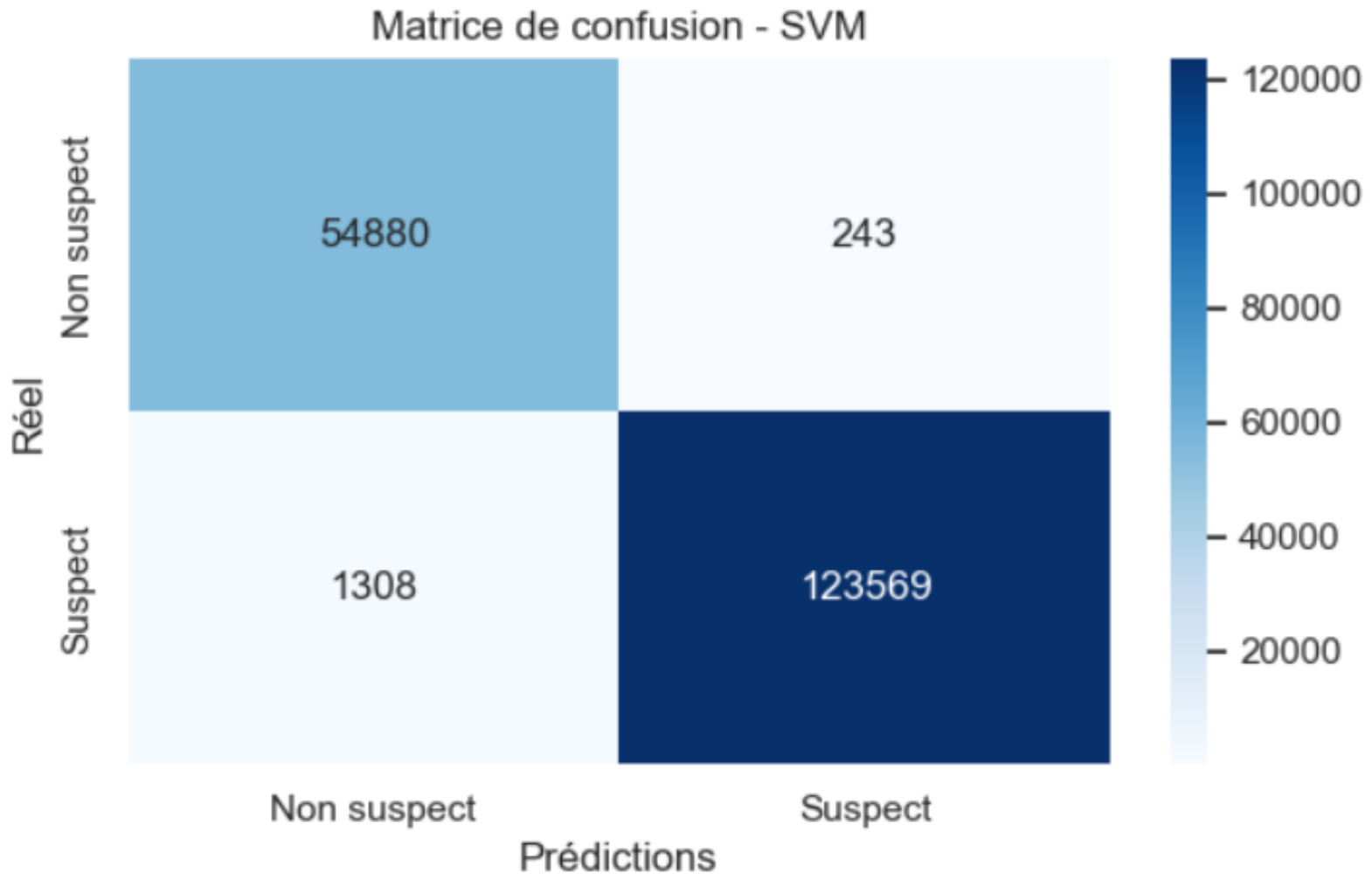
Evaluation du modèle:

- **Accuracy**: Proportion de bonnes prédictions sur l'ensemble des cas
- **Précision:** Parmi les cas prédits comme positifs, combien sont réellement positifs
- **Recall:** Parmi les vrais positifs existants, combien sont bien détectés
- **F1-Score: M**oyenne harmonique entre précision et rappel
- **Matrice de confusion:** Tableau qui résume les prédictions d'un modèle

Interprétation des résultats

- **Accuracy** (Précision globale) : 99.14%
- **Précision** (Précision positive) : 99.80%
- **Recall** (Rappel ou Sensibilité) : 98.95%
- **Matrice de confusion :** nombre raisonnable d'erreurs (1 308 faux négatifs et 243 faux positifs)

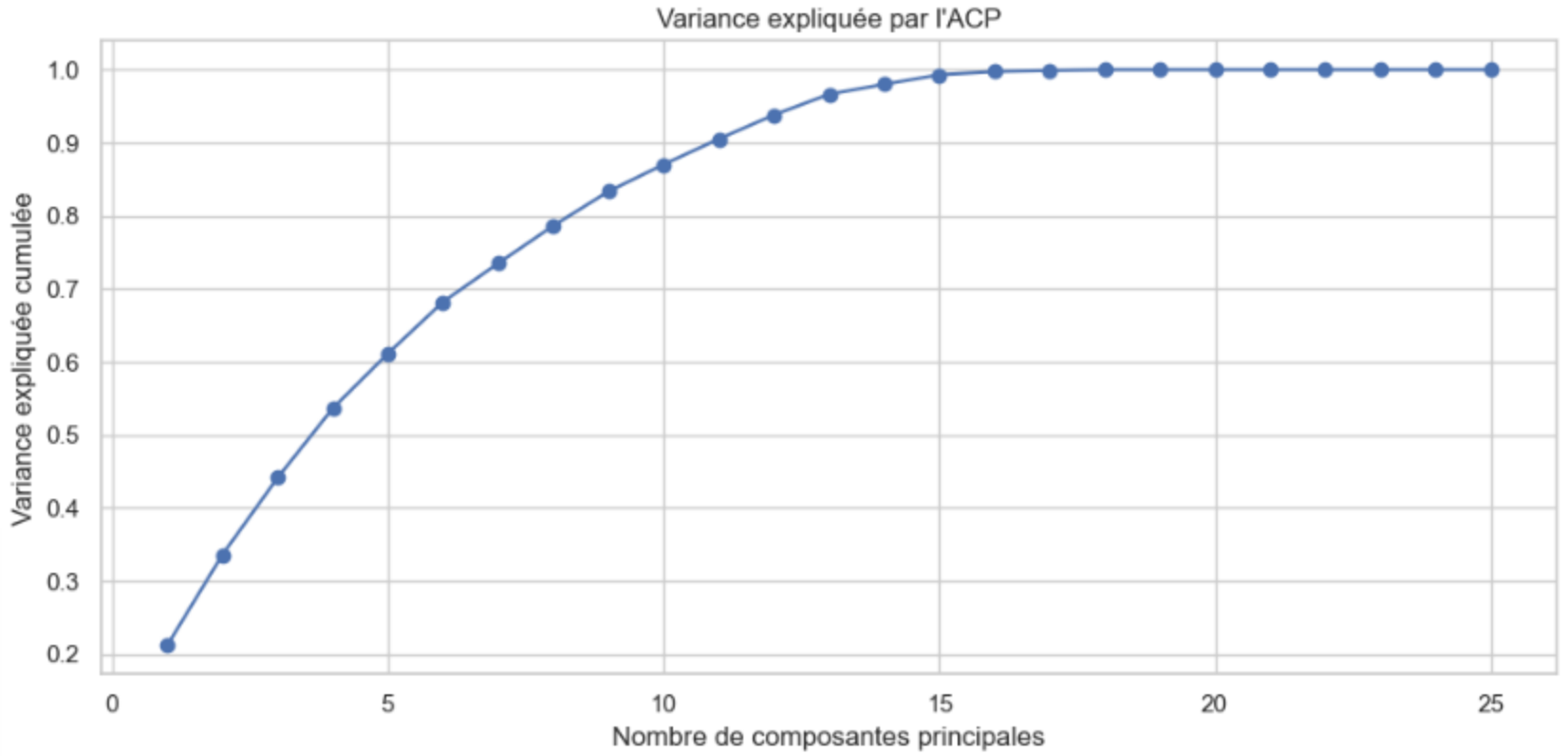
Matrice de confusion



Deuxième approche:

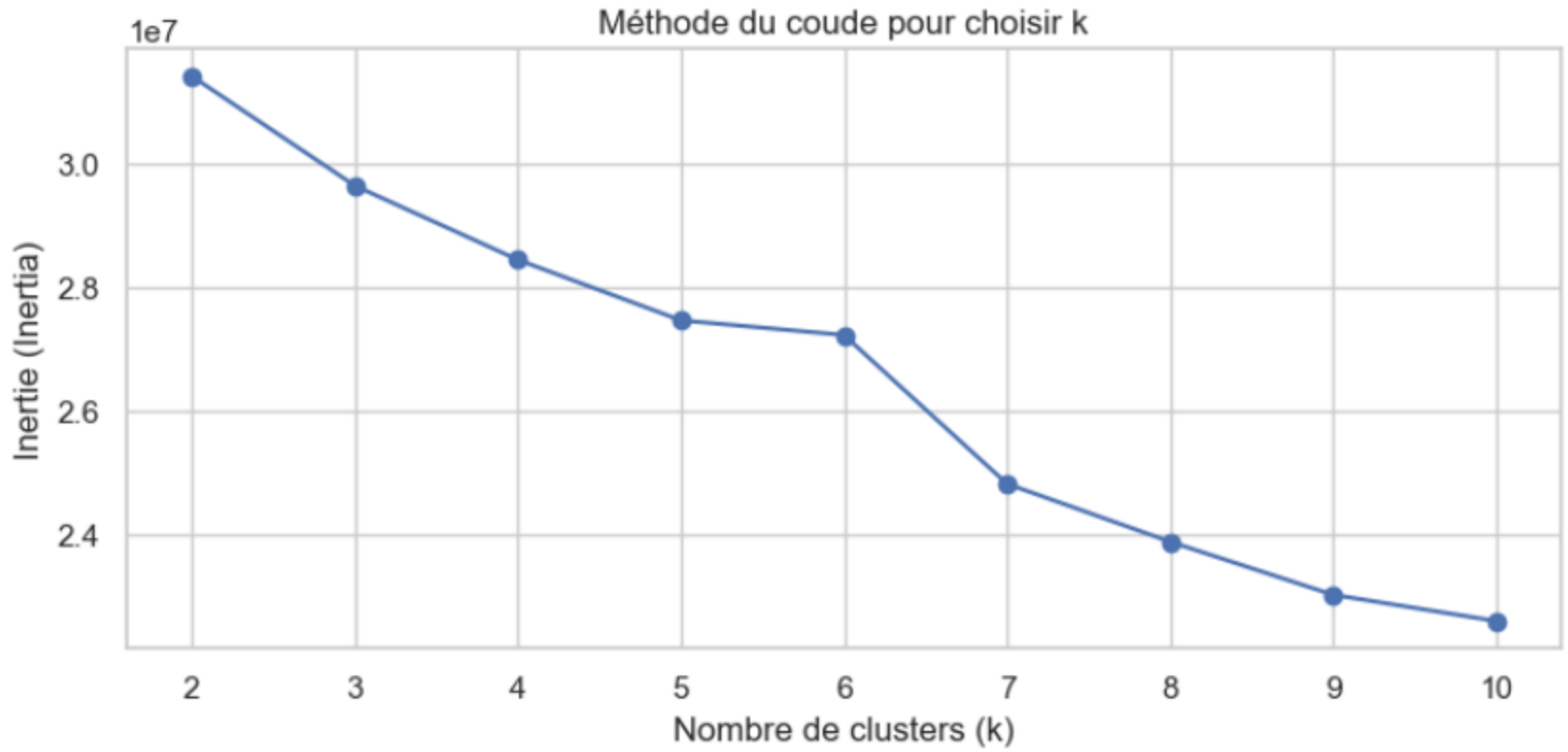
Apprentissage Non supervisé: K Means

Seconde ACP



Nombre de composantes principales à conserver pour expliquer au moins 80% de la variance : 8

Choix du nombre de clusters(k)



Nous retenons donc **k = 6** comme nombre optimal de clusters.

Evaluation du modèle en prenant k = 6

- **Silhouette Score :** Mesure la qualité du regroupement en évaluant à la fois la cohésion au sein d'un cluster et la séparation entre clusters. Un score proche de 1 indique des clusters bien séparés et denses. Un score < 0.5 indique un mauvais regroupement.

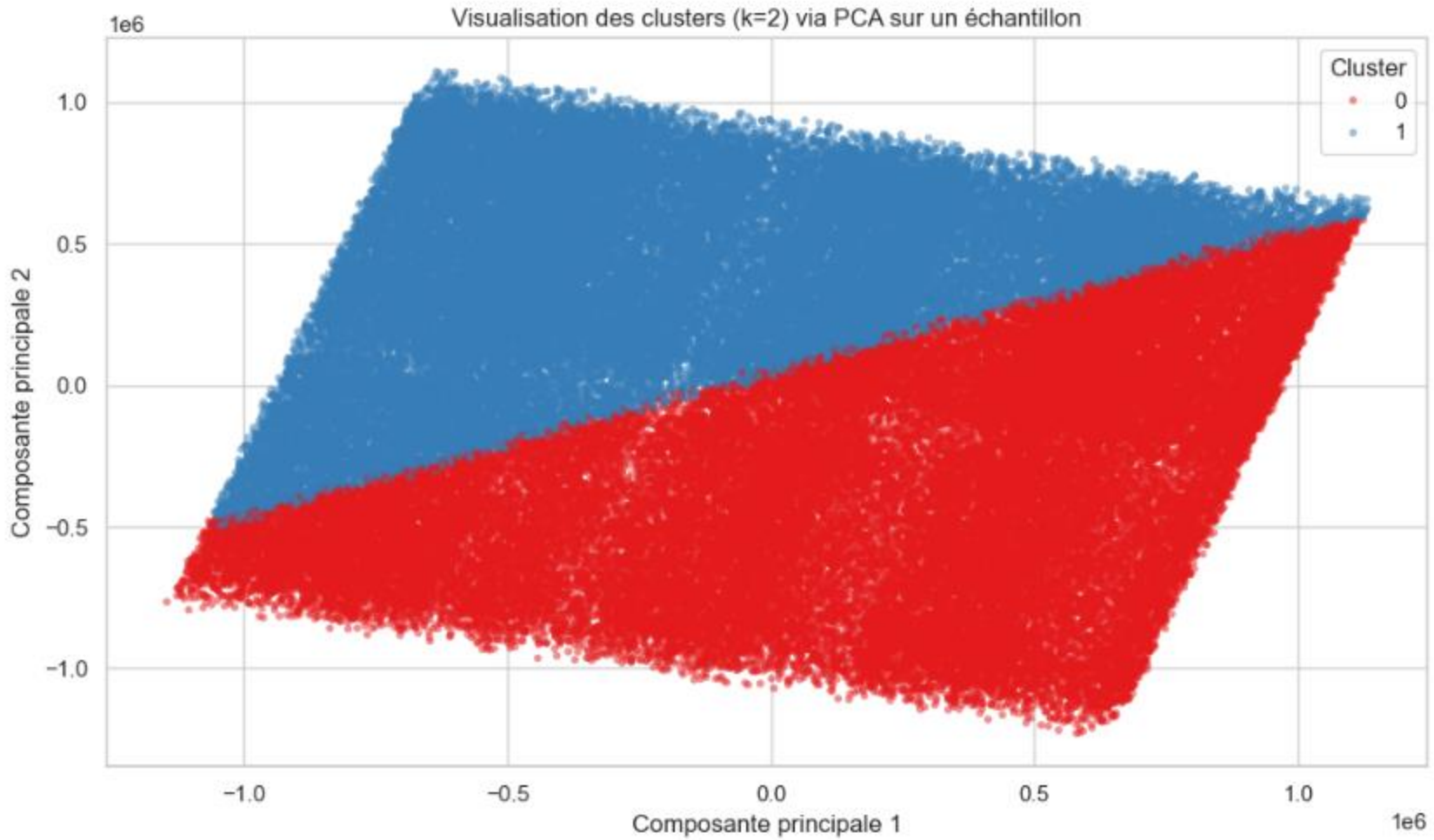
- **Silhouette Score = 0.2906**

Interprétation: Une analyse complémentaire ou une optimisation du nombre de clusters pourrait améliorer la séparation.

Evaluation du modèle en prenant k = 2

- **Silhouette Score** =0.7614
- **Visualisation des clusters (ACP - 2 composantes principales) :** Représentation graphique des utilisateurs projetés sur les deux premières composantes principales issues de l'ACP.

Visualisation des clusters(k=2)



CONCLUSION: COMPARAISON DES RESULTATS

- Les **deux classes sont bien distinctes** dans les données.

- Les **caractéristiques utilisées sont pertinentes** pour différencier les profils.

- Le **clustering est aligné avec la classification**, renforçant la fiabilité de notre analyse.

Merci pour votre attention!!!