Projet SY09 (Printemps 2025) – Rapport final
Pokémon : Qui appartient à la Team Rocket ?

Valentin Ronsseray
valentin.ronsseray@etu.utc.fr

Zhibin Chen
zhibin.chen@etu.utc.fr

Ziad El Hajj Dib
ziad.el-hajj-dib@etu.utc.fr

12/06/2025

1 Introduction

La Team Rocket, organisation criminelle spécialisée dans le vol de Pokémon rares, s'est infiltrée parmi les citoyens de la région de Kanto.

Nous disposons de 4000 individus au statut connu (membres ou non de la Team Rocket) pour entraîner un modèle permettant d'identifier les membres cachés.

Notre objectif : analyser les données pour démasquer les criminels.

2 Analyse exploratoire des données

2.1 Structure générale du jeu de données

Le jeu de données présente une structure riche et variée avec 4000 individus décrits par 17 variables. Cette diversité se traduit par trois catégories principales de variables : 3 variables binaires (`Is Pokemon Champion`, `Rare Item Holder`, `Team Rocket`), 7 variables numériques entières (`Age`, `Average Pokemon Level`, `Criminal Record`, `Win Ratio`, `Number of Gym Badges`, `Number of Migrations` et `Debt to Kanto`), et 7 variables catégorielles (`City`, `Economic Status`, `Profession`, `Most Used Pokemon Type`, `PokéBall Usage`, `Battle Strategy`).

2.2 Analyse de la variable cible `Team Rocket`

L'analyse de la variable d'intérêt révèle un déséquilibre des classes significatif : seulement 18% des individus (720 personnes) appartiennent à la Team Rocket, contre 82% (3280 personnes) qui n'en sont pas membres. Ce déséquilibre constitue un défi méthodologique important pour la modélisation prédictive.

2.3 Comportements discriminants identifiés

L'exploration approfondie des données a révélé des patterns remarquablement nets qui distinguent les membres de la Team Rocket :

2.3.1 Variables quantitatives

L'analyse de la variable `Debt to Kanto` révèle qu'au-delà de 110 000 crédits, tous les individus sont membres de la Team Rocket (figure 4a). Ce seuil est un indicateur fiable pour prédire si un individu appartient ou non à la Team Rocket.

On peut également mentionner le fait que les membres de la Team Rocket présentent également des taux de victoire plus élevés (figure 4b) et une fréquence de migration supérieure (figure 4c), suggérant un mode de vie plus nomade et des compétences de combat développées, cohérents avec leur activité criminelle. Néanmoins, il s'agit d'indicateurs très faibles vis-à-vis de la dette envers Kanto.

2.3.2 Variables binaires

L'analyse révèle quatre propriétés remarquables sur les variables binaires parfaites :

1. Activités caritatives : Tous ceux qui ne participent pas à des œuvres caritatives sont membres de la Team Rocket (figure 4d).
2. Casier judiciaire : Tous les individus ayant un casier judiciaire appartiennent à la Team Rocket (figure 4e).
3. Statut de champion : Tous les champions Pokémon appartiennent à la Team Rocket (figure 4f).

4. Possession d'objets rares : Seuls les membres de la Team Rocket possèdent des objets rares (figure 4g).

2.3.3 Variables catégorielles

Contrairement aux autres types de variables, l'analyse des variables qualitatives (ville, profession, type de Pokémon préféré, etc.) ne révèle aucun pattern discriminant significatif, suggérant que la Team Rocket recrute de manière diversifiée dans toutes les catégories socio-professionnelles et géographiques.

2.3.4 Corrélations entre variables explicatives

La matrice de corrélation des variables quantitatives (figure 4h) montre globalement des corrélations très faibles, à l'exception de deux relations modérées : entre la dette et le ratio de victoires ($= 0,11$), et entre la dette et le nombre de migrations ($= 0,15$). Cette faible intercorrélation est favorable pour éviter les problèmes de multicolinéarité lors de la modélisation.

2.4 Implications pour la prédiction

L'application directe des cinq propriétés identifiées précédemment (cf. sections 2.3.1 et 2.3.2) permet déjà d'atteindre d'excellentes performances globales, sans recourir à un modèle d'apprentissage. En particulier, la précision pour la classe "membre de la Team Rocket" atteint 100%, et le rappel 97.64%, comme le montre le tableau 2.

Ces résultats confirment l'efficacité des variables comportementales et financières pour identifier les criminels. Cependant, bien que le rappel soit élevé, quelques membres de la Team Rocket échappent encore à la détection. C'est précisément pour combler cette marge d'erreur résiduelle que nous poursuivons avec des modèles supervisés. L'objectif est d'explorer si des combinaisons plus fines de variables peuvent permettre d'identifier les cas ambigus, et ainsi d'améliorer encore la couverture de la classe minoritaire.

3 Analyse non supervisée

Afin d'explorer si la structure des données permet d'identifier automatiquement les membres de la Team Rocket, nous avons combiné une analyse en composantes principales (ACP) avec un algorithme de clustering : K-means.

3.1 Analyse en Composantes Principales (ACP)

Nous avons tout d'abord appliqué l'ACP sur l'ensemble des variables après encodage et standardisation. Bien que les deux premières composantes principales n'expliquent qu'environ 7 % de la variance totale, nous avons observé une séparation notable des membres de la Team Rocket le long de la première composante principale (PC1) (figure 5a).

Le *cercle des corrélations* (figure 5b) révèle une structure particulièrement cohérente avec les résultats de notre analyse exploratoire :

— Les variables `Criminal Record`, `Debt to Kanto`, `Rare Item Holder`, `Is Pokemon Champion`, `Number of Migrations` et `Charity Participation` sont toutes fortement corrélées (positivement ou négativement) à PC1 ;
— Or, ces mêmes variables avaient déjà été identifiées comme discriminantes lors de notre analyse descriptive initiale (section 2.3.1 et 2.3.2) ;
— PC1 semble donc synthétiser un **axe comportemental et socialement déviant**, fortement lié à l'appartenance à la Team Rocket ;
— PC2 est quant à elle dominée par les variables `Economic Status_Low` et `Economic Status_Middle`, suggérant une opposition de statut socio-économique.

Ainsi, même si la variance expliquée est faible, la convergence entre les résultats de l'ACP et ceux de l'analyse exploratoire souligne la **valeur structurante** de cette projection. Les composantes principales ne sont pas arbitraires : elles extraient une information directement utile pour la classification.

Pour des raisons de lisibilité, seules les variables dont la norme du vecteur de corrélation est supérieure à 0.2 ont été représentées sur le cercle.

3.2 Clustering avec K-means

Afin de déterminer le nombre optimal de clusters pour l'algorithme K-means, nous avons analysé l'évolution de l'inertie intra-classe (somme des distances au centroïde) pour un nombre de clusters k variant de 1 à 30. La figure 5c présente la courbe obtenue.

La courbe ne présente pas de coude marqué, mais une décroissance rapide de l'inertie est observée jusqu'à $k \approx 5$ ou 6, suivie d'une stabilisation progressive. L'absence de point d'inflexion net suggère que :

— les données ne présentent pas nécessairement de regroupement naturel structuré,

— certaines variables fortement discriminantes (comme *Casier judiciaire* ou *Œuvres caritatives*) influencent fortement la structure des distances,
— les individus sont peut-être répartis selon une distribution continue, sans frontière de cluster claire.

Cependant, notre objectif ici n'est pas de découvrir de nouveaux groupes latents, mais de tester dans quelle mesure un clustering non supervisé peut reproduire la séparation binaire entre *membres de la Team Rocket* et *non-membres*. Pour cela, nous avons volontairement fixé $k = 2$, ce qui correspond à une hypothèse supervisée implicite.

Nous avons ensuite appliqué l'algorithme de k-means (avec $k = 2$) sur deux représentations :
— Sur les variables standardisées, le clustering atteint un **indice de Rand ajusté (ARI)** de 0,842, montrant une forte concordance avec les étiquettes réelles.
— Sur les deux premières composantes principales de l'ACP (figure 5d), l'ARI reste élevé (0,838), ce qui confirme que la réduction dimensionnelle préserve la structure discriminante.

En résumé Malgré une variance expliquée limitée (7%), les deux premières composantes principales capturent des dimensions fortement discriminantes, notamment à travers leur corrélation avec le casier judiciaire, la dette envers Kanto et le nombre de migrations.

L'algorithme de k-means appliqué sur cet espace projeté conserve un ARI élevé (0,838), très proche de celui obtenu sur l'ensemble des variables (0,842), ce qui confirme que l'ACP restitue l'essentiel de la structure utile à la classification. Cela démontre que même en contexte non supervisé, les données présentent une organisation intrinsèque cohérente avec les étiquettes réelles.

4 Apprentissage supervisé

Dans l'objectif de prédire l'appartenance des individus à la Team Rocket, nous avons testé plusieurs algorithmes de classification supervisée sur les 4000 individus étiquetés. La variable cible étant binaire, les modèles évalués incluent : régression logistique, analyse discriminante linéaire (LDA), et k-plus proches voisins (KNN), avec et sans réduction dimensionnelle par analyse en composantes principales (ACP).

4.1 Régression logistique avec ACP

La combinaison de l'ACP (2 composantes) avec une régression logistique offre d'excellentes performances. La validation croisée sur le jeu d'entraînement montre une grande robustesse du modèle, avec une accuracy moyenne proche de 99,4 %.

Une fois entraîné sur l'ensemble des données d'entraînement, le modèle atteint une accuracy de 98,9 % sur le jeu de test indépendant, avec une quasi-parfaite séparation des classes (figure 1).

Ce résultat s'explique notamment par l'existence de propriétés remarquables dans les données (cf. sections 2.3.1 et 2.3.2), qui sont efficacement exploitées par ce modèle linéaire. La projection dans l'espace des composantes principales révèle également une séparation quasi-linéaire entre les individus, ce qui favorise la performance du classificateur.

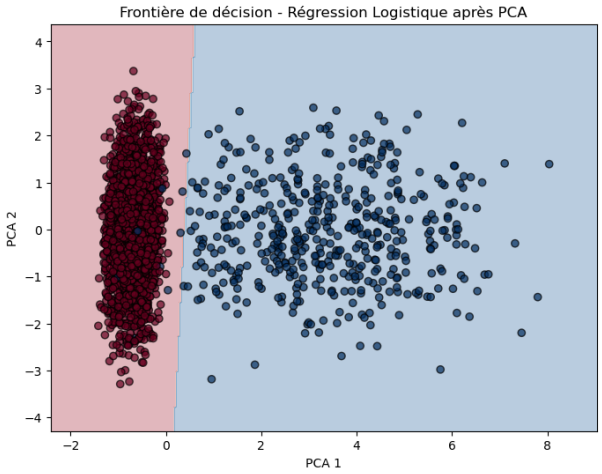


FIGURE 1 – Frontière de décision obtenue avec la régression logistique sur l'ACP (2 composantes)

Le tableau ci-dessous présente les résultats détaillés du rapport de classification obtenu après l'entraînement d'un modèle de régression logistique sur les données projetées dans l'espace des deux premières composantes principales (ACP).

Classe	Précision	Recall	F1-score	Support
0	0.9877	1.0000	0.9938	966
1	1.0000	0.9487	0.9737	234
Accuracy			0.9900	1200
Macro avg	0.9939	0.9744	0.9838	1200
Weighted avg	0.9901	0.9900	0.9899	1200

TABLE 1 – Rapport de classification de la régression logistique après réduction de dimension par ACP

4.2 LDA (Analyse Discriminante Linéaire) avec ACP

L'analyse discriminante linéaire (LDA) a été testée pour la tâche de classification binaire. Ce modèle, adapté à des données où les classes sont séparables de manière linéaire, a donné de bons résultats, avec une accuracy de 97 %.

L'application de la LDA permet une réduction de dimension à une seule composante discriminante (LD1), tout en conservant la capacité de séparer efficacement les deux classes. La projection des individus sur cet axe met en évidence une séparation relativement nette entre les membres de la Team Rocket et les autres individus.

Ce bon résultat s'explique par la présence de propriétés remarquables dans les données, qui sont bien captées par le modèle. Toutefois, une légère confusion reste présente dans les zones proches du seuil critique (notamment en termes de niveau de dette), ce qui peut expliquer les erreurs résiduelles.(figure 2).

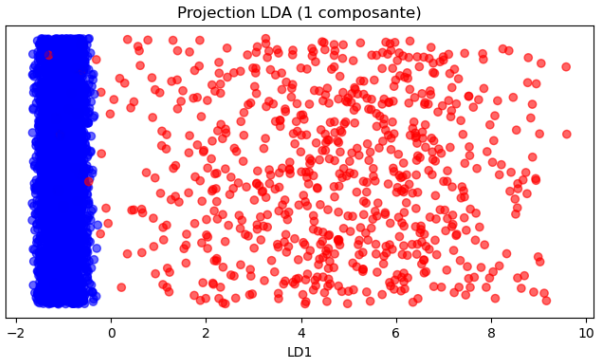


FIGURE 2 – LDA

4.3 KNN (k plus proches voisins) avec ACP

Nous avons ensuite testé l'algorithme des k plus proches voisins (KNN) avec $k = 5$, appliqué après une réduction dimensionnelle des données via l'ACP à 2 composantes. Cette méthode a obtenu une accuracy de 95 %.

Le KNN est un modèle non paramétrique, intuitif mais sensible à la représentation des données. La réduction à deux dimensions a facilité la visualisation de la frontière de décision, mais a aussi entraîné une perte d'information discriminante par rapport à la version initiale. Cela peut expliquer une performance légèrement inférieure par rapport aux autres méthodes testées.

Enfin, le KNN présente une tendance à mal classer les individus situés dans des zones densément mixtes, en particulier lorsque les classes sont proches dans l'espace réduit. L'utilisation d'un nombre différent de voisins ou d'un espace de dimension supérieure pourrait permettre d'améliorer les résultats. (figure 3).

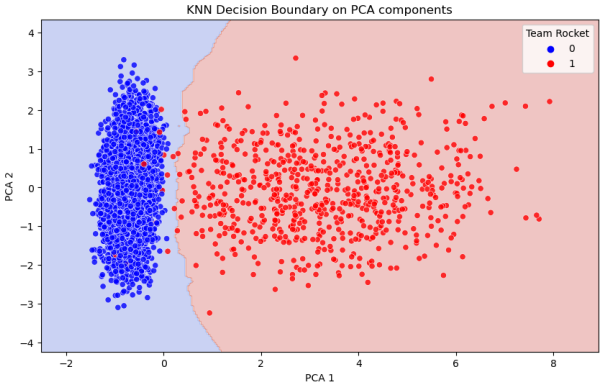


FIGURE 3 – KNN avec ACP

4.4 Conclusion sur la modélisation supervisée

La régression logistique appliquée aux composantes de l'ACP a donné les meilleurs résultats. Cela s'explique par le fait que l'ACP a permis de résumer l'information la plus discriminante, ce qui a facilité la tâche du modèle, bien adapté à une séparation linéaire entre les classes.

5 Conclusion générale

L'objectif de ce projet était de détecter les membres de la Team Rocket à partir d'un ensemble de données mixtes.

L'analyse exploratoire a mis en évidence plusieurs variables fortement discriminantes, dont certaines présentent même des propriétés remarquables permettant une classification quasi-parfaite.

Les méthodes non supervisées, notamment l'ACP combinée au k-means, ont confirmé l'existence d'une structure latente cohérente, même en l'absence d'étiquettes.

Sur le plan supervisé, la régression logistique appliquée aux composantes principales s'est révélée particulièrement efficace, atteignant une précision proche de 99 %. La performance élevée s'explique par la nature très structurée des données, avec un faible recouvrement entre les classes et une forte linéarité des séparations.

En combinant analyse descriptive, réduction dimensionnelle et modèles prédictifs, nous avons montré qu'il est possible d'identifier avec grande précision les membres de la Team Rocket, même dans un contexte de données déséquilibrées et partiellement triviales.

Références

[1] Gérard Govaert, Thierry Denœux, Benjamin Quost, and Stéphane Rousseau. Sy09 - science des données. polycopié de cours. Cours polycopié, UTC, 2025. Université de Technologie de Compiègne.

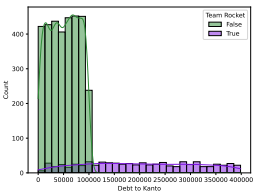
[2] Detective Kotso Kotsopoulos. Pokémon detective challenge : Unmask team rocket. Dépôt de données sur Kaggle, 2025. https://www.kaggle.com/datasets/kotsop/pokmon-detective-challenge.

[3] Benjamin Quost. Projets de sy09 - printemps 2025. Document de projet, UTC, 2025. 18 avril 2025.

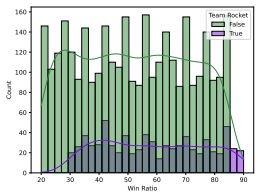
A Annexes

Classe	Précision	Recall	F1-score
Non-membre (0)	0.9948	1.0000	0.9974
Membre (1)	1.0000	0.9764	0.9881
Moyenne (macro)	**0.9974**	**0.9882**	**0.9927**

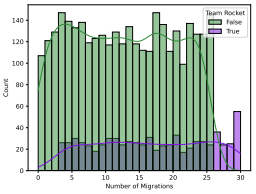
TABLE 2 – Performances de l'application des propriétés remarquables sur l'ensemble des individus étiquetés



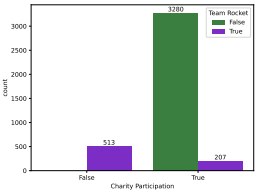
(a) Dette envers Kanto



(b) Ratio de victoires

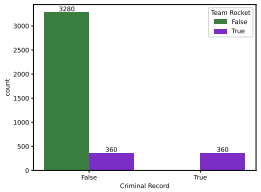


(c) Nombre de migrations

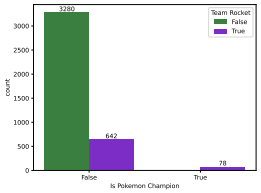


(d) Participation caritative

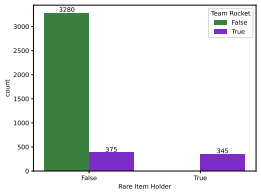
FIGURE 4 – Analyse exploratoire



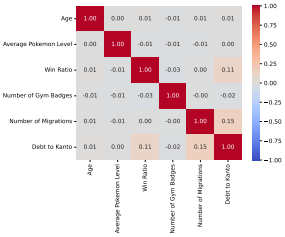
(e) Casier judiciaire



(f) Champions Pokémon

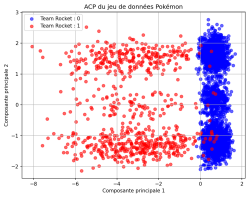


(g) Détenteurs de Pokémon rare

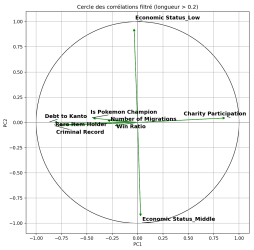


(h) Matrice de corrélation

FIGURE 4 – Analyse exploratoire

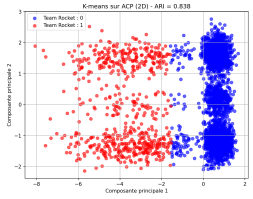


(a) Projection des individus sur les deux premières composantes principales de l'ACP



(b) Cercle des corrélations (ACP), filtré pour les variables les plus contributives

(c) Méthode du coude pour le choix de k



(d) Clustering K-means dans l'espace ACP

FIGURE 5 – Analyse non supervisée