1.Introduction

Ce projet a pour objectif de prédire si un individu appartient à la Team Rocket dans l’univers Pokémon à partir de données issues de Kaggle. Les variables incluent des aspects sociaux, comportementaux et stratégiques.

L’analyse est divisée en deux phases : apprentissage non supervisé (ACP + clustering) et apprentissage supervisé (classification). L’ACP révèle des dimensions fortement liées à l’appartenance à la Team Rocket, telles que le casier judiciaire, les dettes et la participation caritative. Le clustering (K-means, CAH) confirme une structure interne cohérente.

Pour la classification, plusieurs modèles sont testés (KNN, régression logistique, LDA, Random Forest). Le modèle Random Forest obtient les meilleurs résultats avec une précision de 99,5 % et un AUC de 0,99. L’analyse d’importance des variables renforce la compréhension du mécanisme de décision.

Ce travail illustre l’efficacité de la combinaison entre analyses exploratoires et modèles de classification pour résoudre un problème de prédiction complexe.

2.1 Précédente partie consacrée au prétraitement des données et à l'analyse exploratoire

2.2 Analyse en Composantes Principales (ACP)

Afin de comprendre les caractéristiques distinctives entre les membres de la Team Rocket et les autres individus, nous avons appliqué une analyse en composantes principales sur l’ensemble des variables.

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 (PC1), ce qui suggère une structure sous-jacente informative.

Le cercle de corrélation a révélé que les variables suivantes contribuent fortement :

Charity Participation (participation à des œuvres caritatives), fortement positive sur PC1, est typique des non-membres ;

Criminal Record, Debt to Kanto et Rare Item Holder, corrélées négativement avec PC1, sont associées aux membres de la Team Rocket ;

Economic Status\_Low et Middle ont une influence marquée sur la seconde composante (PC2), reflétant une différenciation socio-économique.

Même sans utiliser d’étiquettes supervisées, l’ACP met en évidence des axes d’interprétation très pertinents pour la suite des analyses de classification et de regroupement.

3. Analyse non supervisée (ACP + clustering)

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 deux méthodes de clustering : K-means et classification ascendante hiérarchique (CAH).

Nous avons tout d’abord appliqué l’algorithme de K-means (k=2) sur l’ensemble des variables standardisées. La comparaison entre les clusters obtenus et les vraies étiquettes a donné un indice de Rand ajusté (ARI) de 0,842, ce qui indique une forte concordance. Nous avons ensuite effectué le même clustering sur les deux premières composantes principales, avec un ARI similaire de 0,838, confirmant que l’espace réduit par l’ACP contient une structure discriminante pertinente.

En parallèle, nous avons effectué une CAH sur un échantillon de 500 individus projetés dans l’espace ACP. Le dendrogramme obtenu suggère l’existence de deux groupes bien séparés, cohérents avec les résultats précédents. La CAH présente l’avantage de ne pas nécessiter le choix a priori du nombre de clusters.

En résumé, bien que les deux premières composantes principales n’expliquent qu’environ 7 % de la variance totale, elles permettent une bonne séparation entre membres et non-membres de la Team Rocket. Les résultats de clustering confirment la présence d’une structure sous-jacente exploitable dans les données.

4.1 Analyse supervisée (comparaison des modèles de classification)

Dans le but de prédire si un individu appartient à la Team Rocket, nous avons entraîné cinq modèles supervisés sur les données prétraitées : K plus proches voisins (KNN), régression logistique, analyse discriminante linéaire (LDA), et forêt aléatoire (Random Forest).

Le modèle KNN, avec un nombre optimal de voisins fixé à 3 par validation croisée, atteint une précision de 90 %. Cependant, en raison du déséquilibre des classes, sa capacité à détecter les membres de la Team Rocket est limitée (rappel de seulement 44 %).

La régression logistique et la LDA donnent de très bons résultats. La régression logistique atteint 99 % de précision, un AUC de 0,98, et un rappel de 95 % pour les membres de la Team Rocket. Le modèle LDA se distingue par une excellente précision globale, en particulier pour la classe majoritaire.

Enfin, la forêt aléatoire surpasse tous les autres modèles, avec 99,5 % de précision, 0,99 d’AUC, et seulement quatre erreurs sur les membres de la Team Rocket. L’analyse des importances de variables révèle que les facteurs “Debt to Kanto”, “Charity Participation” et “Criminal Record” sont les plus discriminants, ce qui corrobore parfaitement les résultats obtenus lors de l’analyse en composantes principales.

Ainsi, nous avons choisi d’adopter le modèle Random Forest comme méthode principale de classification pour la suite de notre projet.

4.2 Ajustement des seuils de classification

Dans le modèle de forêt aléatoire, nous avons quatre erreurs de classification (des membres de l'équipe Rocket ont été considérés comme des personnes ordinaires), alors que la classification des personnes ordinaires est correcte à 100 %. Nous avons donc décidé d'adopter une stratégie plus agressive en abaissant le seuil de décision pour déterminer s'il s'agit d'un membre de l'équipe Rocket, afin de rendre la classification plus sensible. Lorsque le seuil est de 0,25, il n'y a que deux erreurs de classification, et le taux de précision passe de 99,5 % à 99,75 %.

Nous avons analysé les échantillons qui ont encore été mal classés. Leur probabilité de prédiction était au minimum de 0,00, mais il s'agissait bien de membres de l'équipe Rocket.

Par rapport aux autres échantillons, ceux-ci étaient très similaires aux personnes ordinaires dans l'espace des caractéristiques, et nous n'avons pas pu les distinguer correctement même en abaissant le seuil de classification à 0,25. Cette erreur de classification peut être due à une erreur dans les étiquettes elles-mêmes ou au fait qu'il s'agit de véritables membres « bien déguisés »

5.Conclusion

À l’issue de ce projet, nous avons pu prédire avec une grande précision l’appartenance à la Team Rocket grâce à une approche méthodique combinant analyse exploratoire et modèles de classification. Le modèle Random Forest s’est révélé le plus performant, tout en offrant une bonne interprétabilité via l’analyse des variables.

Les résultats du clustering et de l’ACP corroborent les performances supervisées, soulignant une structure interne forte et exploitable dans les données.

Des pistes d’amélioration incluent l’enrichissement des variables sociales, l’utilisation de modèles plus complexes (deep learning) et l’évaluation de la robustesse sur des jeux de données alternatifs.