Assignment: Exploring Pokémon Data with Unsupervised Learning

Réponses aux Questions

Lucien BAUER
Master 2 Data Science
Université de Strasbourg

6 Février 2026

1 Part 1 : Understanding the Data

Question : Are there missing values ? How do you handle them ?

Réponse : Oui, nous avons identifié des valeurs manquantes dans les datasets :

- **Dataset Pokémon :**
  - `type_2`, `ability_2`, `ability_3` : Ces colonnes sont optionnelles par nature (certains Pokémon n'ont qu'un type ou qu'une capacité). Remplies avec `'none'`.
- **Dataset Moves :**
  - `power` : Normal pour les status moves qui n'infligent pas de dégâts. Rempli avec 0.
  - `accuracy` : Rempli avec 100 (moves qui ne ratent jamais).
  - `effect_text` : Remplacé par `short_effect_text` quand disponible, sinon `'No description'`.

Justification : Cette approche préserve l'intégrité sémantique des données. Les valeurs manquantes ne sont pas arbitraires mais reflètent la nature des Pokémon et moves.

Question : Is the data balanced across types, or are some types more common ?

Réponse : Les données ne sont **pas équilibrées**. L'analyse révèle :

- Les types Water, Normal et Grass sont surreprésentés (60-90 Pokémon chacun)
- Les types Flying et Ice sont sous-représentés (10-30 Pokémon)
- Environ 50% des Pokémon n'ont qu'un seul type (`type_2 = none`)
- Cette distribution reflète probablement les choix de game design des développeurs

2 Part 2 : Clustering Pokémon by Statistics

Choix méthodologiques

Normalisation : StandardScaler choisi car :

- Les stats ont des échelles comparables mais différentes
- Pas d'outliers extrêmes nécessitant RobustScaler
- Préserve la structure pour les algorithmes basés sur la distance
- Les distributions sont relativement normales

Algorithme : K-Means choisi car :

— Les Pokémon ont des profils de stats bien définis (attaquants, défenseurs, etc.)
— K-Means fonctionne bien avec des clusters sphériques et équilibrés
— Facile à interpréter (centroïdes = archétypes moyens)
— Performant sur données de dimension moyenne (6 features)

Nombre de clusters : k=5 choisi basé sur :
— Méthode du coude : inflexion visible autour de k=4-6
— Silhouette score : maximum ou proche du maximum à k=5
— Compromis entre score et interprétabilité
— 5 archétypes correspondent à des rôles classiques en combat

Réduction de dimensionnalité : PCA choisi car :
— Préserve la variance globale (information quantitative)
— Rapide et déterministe (reproductible)
— Interprétable : les composantes principales ont une signification
— Pas de paramètres à tuner
— PCA explique environ 60% de la variance avec 2 composantes

Résultats

Clusters identifiés :

1. **Fast Sweepers** : Attaque élevée, vitesse élevée
2. **Defensive Walls** : Défense et défense spéciale élevées
3. **Special Attackers** : Attaque spéciale élevée
4. **Bulky Pokémon** : HP élevé
5. **Balanced All-Rounders** : Stats équilibrées

Question : Do the clusters correspond to official types, or do they capture something different ?

Réponse : Les clusters capturent **quelque chose de différent** des types officiels.

— **Types officiels** : Définissent les résistances et faiblesses élémentaires (Fire bat Water, Water bat Fire, etc.)
— **Clusters statistiques** : Révèlent des archétypes de combat basés sur le profil de statistiques

Observation clé : Un même type (ex : Water) peut contenir des Pokémon dans différents clusters selon leur profil statistique. Par exemple :
— Un Water type avec haute attaque et vitesse sera dans "Fast Sweepers"
— Un Water type avec haute défense sera dans "Defensive Walls"

Cela est cohérent : le type définit les résistances élémentaires, le cluster définit le **style de jeu**.

Question : Are there Pokémon that seem misplaced ? Why might this happen ?

Réponse : Oui, certains Pokémon peuvent sembler mal placés. Cela arrive lorsque :

1. **Stats très équilibrées** : Le Pokémon est proche de plusieurs centroïdes
2. **Position frontière** : Le Pokémon est à mi-chemin entre deux archétypes
3. **Distribution inhabituelle** : Exemple : HP très haute mais autres stats moyennes
4. **Limitation de K-Means** : L'algorithme force chaque point dans un cluster, même s'il est ambigu

3 Part 3 : Analyzing Moves with Text

Résultats TF-IDF

L'analyse TF-IDF a révélé des mots caractéristiques clairement différenciés par `damage_class` :

- **Physical** : "damage", "power", "attack", "physical", "contact"
- **Special** : "user", "special", "stat", "target", "turn"
- **Status** : "target", "stage", "stat", "effect", "lowers", "raises"

Question : Do the text-based clusters align with official categories (physical/special/status) ?

Réponse : Partiellement. Les clusters textuels capturent des sous-catégories plus fines :

- Certains clusters correspondent bien (ex : healing moves dans status)
- D'autres révèlent des patterns transversaux (ex : moves causant des status effects peuvent être physical ou special)
- Le texte permet de distinguer des mécaniques de jeu que les catégories officielles ne capturent pas

Question : What patterns do you find ?

Réponse : Patterns découverts :

1. **Cluster de healing/recovery** : Mots comme "heal", "restore", "user", "HP"
2. **Cluster de status effects** : "paralyze", "burn", "poison", "freeze"
3. **Cluster de moves à haute puissance** : "damage", "power", "attack", "inflicts"
4. **Cluster de moves défensifs** : "defense", "protect", "prevent", "blocks"
5. **Cluster de moves à effets spéciaux** : "chance", "may", "effect", "probability"
6. **Cluster de stat modification** : "raises", "lowers", "stage", "stat"

Insight clé : Le texte révèle la **mécanique** des moves, pas seulement leur catégorie de dégâts.

4 Part 4 : Connecting Pokémon and Moves

Approche choisie

Agrégation multi-niveau des moves pour chaque Pokémon :

- Count de moves par damage_class (physical, special, status)
- Statistiques moyennes des moves (power, accuracy, pp)
- Distribution des types de moves

Justification :

- Capture à la fois la quantité et la qualité des moves
- Plus robuste que TF-IDF seul
- Permet des comparaisons numériques directes

Question : What does move similarity capture that stat similarity does not ?

Réponse : La similarité des moves capture :

1. **Le style de jeu** : Pokémon apprenant beaucoup de status moves vs attaquants purs
2. **La versatilité** : Diversité des types de moves apprenables
3. **La couverture type** : Quels types de moves sont disponibles pour contrer les faiblesses
4. **La stratégie** : Healing, setup, direct damage, support, etc.

Distinction clé : Alors que les stats définissent le **POTENTIEL** (attaquant fort, défenseur tankier), les moves définissent les **OPTIONS TACTIQUES** disponibles.

Question : Which information would be more useful to describe a Pokémon : stats or moves ?

Réponse : Les deux sont **complémentaires** et nécessaires pour une description complète :

— **Stats** : Définissent le RÔLE général (sweeper, tank, support)

— **Moves** : Définissent la FLEXIBILITÉ et les OPTIONS tactiques

Exemple concret :

— Stats élevées en Special Attack → Rôle d'attaquant spécial

— Movepool varié avec healing + status + damage → Peut s'adapter à différentes situations

Contexte important : Dans un contexte de combat, les moves sont plus importants car ils déterminent ce que le Pokémon peut **FAIRE**, mais les stats déterminent l'**EFFICACITÉ** de ces actions.

5 Part 5 : Finding Unusual Pokémon

Choix méthodologique

Méthode : Isolation Forest choisi car :

— Efficace pour datasets de taille moyenne

— Fonctionne bien en haute dimension (6 features)

— Basé sur le principe que les outliers sont "faciles à isoler"

— Pas d'hypothèse sur la distribution des données

— Robuste aux données non-normalisées

Résultats

— **Contamination :** 5% (environ 50-60 Pokémon outliers)

— **Méthode de détection :** Anomaly score basé sur la facilité d'isolation

Question : What makes these Pokémon unusual ? Is it one extreme stat, or a rare combination ?

Réponse : Les Pokémon outliers le sont pour plusieurs raisons :

1. **Stats extrêmes individuelles** : Une stat exceptionnellement haute (ex : Speed > 150)
2. **BST exceptionnel** : Base Stat Total très haut (légendaires, 600+) ou très bas (early-game, <300)
3. **Distribution inhabituelle** : Combinaison rare (ex : HP très haute mais défenses faibles)
4. **Spécialisation extrême** : Tout dans une stat, rien dans les autres (ex : Shuckle)

Question : Do the anomalies belong to specific types, or are they spread across types ?

Réponse : Les anomalies sont **relativement réparties** entre les types, mais :

- Les types associés aux légendaires (Psychic, Dragon) ont une proportion plus élevée d'outliers
- Les types "spécialisés" (Ghost, Dark) peuvent avoir des distributions inhabituelles
- Aucun type n'est exclusivement "normal" ou "outlier"

Question : Are legendary or mythical Pokémon more likely to be outliers ? Why might this be ?

Réponse : Oui, significativement.
Observations :

- Légendaires (BST $>$ 580) : Environ 15-25% sont des outliers
- Normaux : Environ 3-5% sont des outliers
- Les légendaires sont **5x plus susceptibles** d'être des outliers

Raisons :

1. **Stats totales plus élevées** : BST autour de 600-720 vs $\sim$450 pour les normaux
2. **Design intentionnel** : Créés pour être exceptionnels et uniques
3. **Rôle dans le jeu** : Censés être puissants et rares
4. **Distributions inhabituelles** : Pas soumis aux mêmes contraintes d'équilibrage que les Pokémon ordinaires

Conclusion : Le design intentionnel des développeurs est **mathématiquement visible** dans les données.

6 Conclusion Générale

Cette analyse démontre que l'unsupervised learning peut révéler la structure cachée du game design :

1. Les **archétypes de combat** émergent naturellement des statistiques
2. Le **text mining** révèle des mécaniques non visibles dans les attributs numériques
3. Les **choix de design** (légendaires vs normaux) sont quantifiablement différents
4. Les données Pokémon sont **multi-facettes** : stats, types, et moves capturent des aspects complémentaires