Assignment: Exploring Pokémon Data with Unsupervised Learning

January 23, 2026

General Information

Work mode: Individual or groups (max 3 students)
Deadline: February 6, 2025 (presentation in class)
Submission: Jupyter notebook (.ipynb)

On February 6, we will have presentations in class. Each student or group will briefly present their work (5–10 minutes). Be ready to explain your choices and discuss your main results (or what you found interesting :)).

1 Introduction

You have three datasets describing Pokémon, their moves, and which moves each Pokémon can learn. Your goal is to use unsupervised learning techniques to discover patterns and answer questions about the structure of this data.

This assignment is not about finding "the best Pokémon." It is about discovering patterns, measuring similarity, and understanding what different data sources tell us.

2 Datasets

You are provided with three CSV files[1]:

`pokemon_complete.csv`

Each row is a Pokémon with the following columns:

[1] Data were retrieved from https://pokeapi.co/

Column	Description
pokemon_id	Unique identifier
name	Pokémon name
height	Height (in decimeters)
weight	Weight (in hectograms)
base_experience	Experience points gained when defeated
type_1	Primary type (grass, fire, water, etc.)
type_2	Secondary type (can be empty)
ability_1, ability_2, ability_3	Abilities (can be empty)
hp	Health Points
attack	Physical attack strength
defense	Physical defense
special-attack	Special attack strength
special-defense	Special defense
speed	Speed in battle

moves_complete.csv

Each row is a move with the following columns:

Column	Description
move_id	Unique identifier
name	Move name
type	Move type (normal, fire, water, etc.)
power	Base power (can be empty for status moves)
pp	Power Points (how many times the move can be used)
accuracy	Accuracy percentage
priority	Move priority in battle order
damage_class	Category: physical, special, or status
effect_text	Full text description of the move effect
short_effect_text	Short description

learnset_complete.csv

Each row links a Pokémon to a move it can learn:

Column	Description
pokemon_id	Pokémon identifier (links to pokemon.csv)
move_id	Move identifier (links to moves.csv)
move_name	Move name

3 Part 1: Understanding the Data

Before applying any technique, explore the data.

Tasks:

1. Load the three datasets and check their structure.
2. Report basic statistics:
    - How many Pokémon are there?

- How many moves are there?
- How many moves per Pokémon on average?

3. Show the distribution of Pokémon types and move categories (damage_class).

Questions to answer:

- Are there missing values? How do you handle them?
- Is the data balanced across types, or are some types more common?

4 Part 2: Clustering Pokémon by Statistics

Main question: Do Pokémon naturally group into distinct "archetypes" based on their statistics?

Tasks

2.1 Prepare the data

- Select the statistics: `hp`, `attack`, `defense`, `special-attack`, `special-defense`, `speed`
- Normalize the data and explain your choice (StandardScaler, MinMaxScaler, or other)

2.2 Apply clustering

- Choose a clustering algorithm (K-means, hierarchical clustering, DBSCAN, or another method)
- Explain why you chose this algorithm for this data
- If your algorithm requires parameters (e.g., number of clusters, epsilon), test different values and justify your final choice
- Use an appropriate metric (e.g., silhouette score, dendrogram, elbow method) to support your decision

2.3 Visualize the clusters

- Apply a dimensionality reduction technique (PCA, UMAP, t-SNE, or another method) to reduce the data to 2 dimensions
- Briefly explain why you chose this technique
- Create a scatter plot colored by cluster
- Create another scatter plot colored by primary type (for comparison)

2.4 Interpret the clusters

- For each cluster, compute the average statistics
- Give each cluster a descriptive name (e.g., "fast attackers," "defensive tanks")
- Show 3–5 example Pokémon from each cluster

Questions to answer:

- Do the clusters correspond to official types, or do they capture something different?
- Are there Pokémon that seem misplaced? Why might this happen?

5 Part 3: Analyzing Moves with Text

Main question: What can we learn from the text descriptions of moves that we cannot see from the numeric attributes?

Tasks

3.1 Apply TF-IDF to move descriptions

- Use the `effect_text` column
- Preprocess the text (lowercase, remove punctuation)
- Compute TF-IDF vectors for all moves

3.2 Find characteristic words

- For each damage class (physical, special, status), find the 10 words with highest average TF-IDF
- Display these in a table

3.3 Cluster the moves

- Apply K-means to the TF-IDF vectors
- Choose an appropriate number of clusters
- Examine what moves end up in the same cluster

Questions to answer:

- Do the text-based clusters align with the official categories (physical/special/status)?
- What patterns do you find? (e.g., healing moves, moves that cause status effects, high-damage moves)

6 Part 4: Connecting Pokémon and Moves

Main question: Are Pokémon that are similar in statistics also similar in the moves they can learn?

Tasks

4.1 Create a move-based representation of Pokémon

For each Pokémon, aggregate information about the moves it can learn. You can use one of these approaches (or propose your own):

- Count of moves per damage class (physical, special, status)
- Average move power, accuracy, etc.
- TF-IDF of combined move descriptions

Document and justify your choice.

4.2 Compare similarity structures

- Pick 5 Pokémon of your choice
- For each, find the 3 most similar Pokémon using **statistics only**

- For each, find the 3 most similar Pokémon using **move information only**
- Present results in a table

4.3 Analyze the relationship

Find at least one example of:

- Pokémon similar in stats AND similar in moves
- Pokémon similar in stats BUT different in moves
- Pokémon different in stats BUT similar in moves

Questions to answer:

- What does move similarity capture that stat similarity does not?
- Which information would be more useful to describe a Pokémon: stats or moves?

7 Part 5: Finding Unusual Pokémon

Main question: Are there Pokémon that are unusual or do not fit well with the others?

Tasks

5.1 Apply anomaly detection

- Choose an anomaly detection method (e.g., Isolation Forest, LOF, DBSCAN outliers, distance from cluster centroids, or another approach)
- Apply it to the same data you used for clustering
- Identify the top 5–10 most anomalous Pokémon and try to understand *why* it is considered an outlier

Questions to answer:

- What makes these Pokémon unusual? Is it one extreme stat, or a rare combination?
- Do the anomalies belong to specific types, or are they spread across types?
- Are legendary or mythical Pokémon more likely to be outliers? Why might this be?

8 Practical Tips

- **Start simple.** Get basic clustering working before trying complex representations.
- **Document your choices.** There is no single correct answer, but you must explain your reasoning.
- **Look at examples.** Numbers and plots are useful, but examining individual Pokémon helps build intuition.
- **Unexpected results are interesting.** If something does not work as expected, discuss why.

Optional Bonus

Design a method to select a "balanced team" of 6 Pokémon. Define what "balanced" means (diversity? type coverage? complementary roles?) and implement a selection procedure.

This is deliberately open-ended. A simple, well-justified approach is better than a complex one you cannot explain.

Good luck! See you on February 6 for the presentations.