Pokémon Learning: Clustering applicato al mondo dei Pokémon

Justin Armanini, Marco Donzella, Francesco Spampinato, Matteo Tarli

Indice

1 Abstract

L'obiettivo di questa ricerca è capire se esistono gruppi di Pokémon sulla base delle loro caratteristiche native. E' stato verificato che in rete è disponibile un'ampia letteratura per rispondere a problemi di classificazione dei Pokémon, per tal motivo si è deciso di focalizzare questa ricerca sulla possibilità di utilizzare la Cluster Analysis. E' stata effettuata una prima analisi considerando gli attributi di battaglia (numerici), ed una seconda analisi inerente gli attributi relativi le caratteristiche fisiologiche (numerici e categorici). Per quanto riguarda la prima domanda di ricerca, si è giunti a distinguere quattro gruppi di Pokèmon tramite l'algoritmo K-means, mentre per la seconda sono stati identificati sei prototipi di riferimento con l'algoritmo K-medoids.

2 Introduzione

"Pokémon" è un media franchise giapponese di proprietà di *The Pokémon Company* creato nel 1996 da Satoshi Tajiri, il quale ha dato vita ad una serie animata, ad una serie di videogiochi realizzati in forma Role Playing Game (RPG) pubblicati da Nintendo e ad un gioco di carte collezionabili. I Pokémon sono delle creature immaginarie: lo scopo del gioco è quello di catturarli, allenarli e farli sfidare tra loro in battaglie virtuali. La cattura, che è una componente fondamentale del gioco, avviene utilizzando degli oggetti chiamati "Poké Ball". Nel videogioco sono presenti diversi tipi di Poké Ball che differiscono per la loro efficacia. Questa dinamica incita il videogiocatore a competere per la caccia ai Pokémon più forti. Tra questi spiccano sicuramente i cosiddetti Pokémon "leggendari": essi sono dei Pokémon rari, quindi molto difficili da trovare e catturare. Nel videogioco, ogni Pokémon è caratterizzato da degli attributi di battaglia che ne determinano il rendimento nelle sfide (ad esempio attacco, difesa, velocità) e da caratteristiche fisiologiche che descrivono la natura del Pokémon (ad esempio altezza, peso). Lo scopo di questa ricerca è individuare eventuali raggruppamenti di Pokèmon sulla base dei loro attributi a livello iniziale. Pertanto, si è tentato di rispondere alle seguenti domande attraverso la Cluster Analysis:

1. È possibile suddividere i Pokémon in gruppi, considerando le loro caratteristiche di battaglia?

2. E' possibile suddividere i Pokémon in gruppi, considerando unicamente le loro caratteristiche fisiche?

3 Esplorazione dei dati

Il dataset in considerazione (*pokemon_alopez247*) consiste di 721 osservazioni e di 23 caratteristiche per ciascuna osservazione. Il dataset contiene tutti i Pokémon dalla prima generazione(1996-1999) fino alla sesta (2013–2016). Sono presenti sia variabili riguardanti le caratteristiche di battaglia di un Pokémon (e.g. *Hp*, *Attack*), sia caratteristiche fisiologiche (e.g. *Weight_KG*, *Type_1*,*Body_Style*). La variabile *Total* è data dalla combinazione lineare semplice delle variabili di battaglia. Nel dataset sono inoltre presenti i Pokémon leggendari (46),spesso contraddistinti da statistiche di battaglia rilevanti. Per questo motivo è possibile ipotizzare la presenza di valori anomali nell'analisi, che dovranno essere esclusi. Di seguito si riporta il grafico delle frequenze assolute delle modalità del carattere *Is_Legendary*.

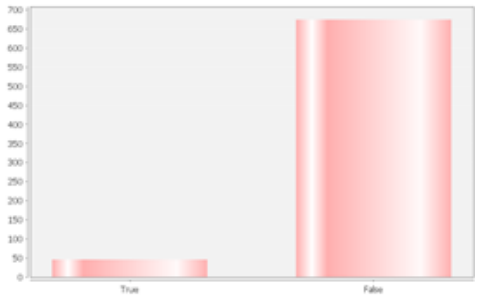

Figura 1: Frequenza Pokemon Rari

Inoltre nel dataset sono presenti alcuni valori mancanti nella variabili *Type_2*, *Pr_Male* e *Egg_Group2*. Questo è dovuto al fatto che molti Pokemon non possiedono queste determinate caratteristiche. Infine si è osservato che è presente una correlazione di circa -0.75 tra la variabile *Total* e la variabile *Catch_Rate*, questo perché tendenzialmente i Pokémon più forti sono anche i più difficili da catturare. La variabile *Catch_Rate* infatti permette di calcolare la probabilità di catturare il Pokémon nel seguente modo:

$$p_{cattura} = \frac{Stato}{Ball+1} + \frac{Tasso+1}{Ball+1} * \frac{f+1}{256}$$

Dove

- *Stato* vale 25 se il Pokémon è addormentato o congelato, 12 se è paralizzato, scottato o avvelenato e 0 altrimenti.
- *Ball* vale 255 per la Poké Ball, 200 per la Mega Ball e 150 per le altre Ball (Ultra Ball e Safari Ball).
- f è una funzione che considera la Ball utilizzata e gli *Hp* correnti.

4 Clustering caratteristiche di battaglia

Per rispondere alla prima domanda di ricerca si è proseguito conducendo una Cluster Analysis usando i seguenti metodi: K-means, K-medoids, Hierarchical Average e Hierarchical Ward.

4.1 Preprocessing

Per condurre questo primo tipo di analisi, sono state selezionate le variabili riguardanti solamente le caratteristiche di battaglia native dei Pokémon:

- *Hp*: Hit Points, i punti salute del Pokémon
- *Attack*: determina il valore di danno che un pokémon infligge quando utilizza mosse fisiche.
- *Defense*: determina il valore di danno cui il Pokémon riesce a resistere quando subisce una mossa fisica.
- *Sp_Atack*: determina il valore di danno che un pokémon infligge quando utilizza mosse speciali.
- *Sp_Defense*: determina il valore di danno cui il Pokémon riesce a resistere quando subisce una mossa speciale.
- *Speed*: determina quale Pokemon in battaglia sarà il primo ad agire.

La variabile *Total* è stata esclusa dall'analisi in quanto, come osservato in fase di esplorazione, non avrebbe aggiunto informazioni. Per verificare l'eventuale presenza di multicollinearità è stata calcolata la matrice delle correlazioni per le suddette variabili:

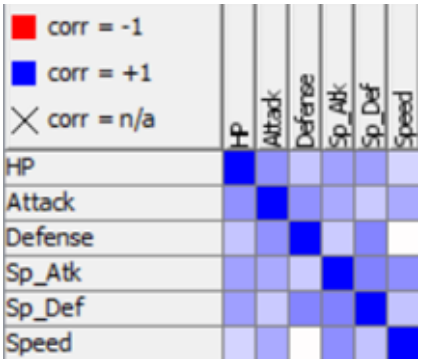

Figura 2: Matrice di correlazione

Come si evince dalla visualizzazione sopra riportata, tutte le variabili, in maniera più o meno forte, risultano essere correlate positivamente (tra 0 e 0.5). Ne consegue che nel mondo dei Pokémon non si verifica il cosiddetto fenomeno del "Glass Cannon" tipico dei videogiochi di combattimento (come ad esempio Final Fantasy, League of Legends e Dungeons & Dragons), secondo cui i personaggi che spiccano per caratteristiche offensive, al tempo stesso manifestano debolezze per quelle difensive.Per questo vale la pena investigare quale siano eventuali logiche che permettono di raggruppare i Pokemon. Non essendo presenti dei valori mancanti nelle caratteristiche di battaglia, non è stato necessario alcun intervento in questo ambito. Tuttavia, come affermato nell'introduzione, nel dataset sono presenti i Pokémon leggendari, che potrebbero sbilanciare le distribuzioni delle variabili considerate. Pertanto si procede con la rimozione dei record con valori degli attributi non appartenenti agli intervalli $(Q1_i - 3 * IQR_i, Q3_i + 3 * IQR_i)$ con i=1,2,..,6 delle sei variabili considerate. Di seguito si riportano due esempi di Pokémon rimossi.

Figura 3: Shuckle

Figura 4: Regirock

Shuckle è un pokémon di seconda Generazione; risulta essere il più sbilanciato in quanto caratterizzato da valori di attacco e HP estremamente bassi, ma dalla migliore difesa in assoluto. Regirock è un Pokémon leggendario di terza generazione ed è caratterizzato da elevata Difesa, pari a 200.

Prima di procedere alla clusterizzazione, avendo le variabili campo di variazione differente, si è optato per normalizzare i dati riscalando sull'intervallo $[0, 1]$. Come metodo di valutazione per individuare il numero ottimale di cluster si è deciso di studiare l'andamento del coefficiente di Silhouette al variare del numero di cluster. Essendo lo studio concentrato sull'individuazione di gruppi, si è deciso di indirizzare il lavoro su un numero di clusters superiore a due per evitare un semplice raggruppamento in forti e deboli, che non risulterebbe interessante.

4.2 K-means

Il primo algoritmo di clustering che è stato utilizzato è K-means, utilizzando la distanza euclidea come misura di dissimilarità. Osservando l'andamento dell'indice di Silhouette, l'algoritmo K-Means è stato implementato usando k=4 come numero di cluster ottimale.

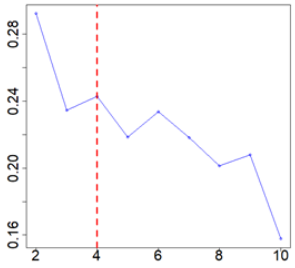

Figura 5: Silhouette K-means

Di seguito si riportano i centroidi dei cluster ottenuti:

Row ID	HP	Attack	Defense	Sp_Atk	Sp_Def	Speed
cluster_0	50.018	53.396	51.143	47.689	48.879	49.392
cluster_1	71.988	82.171	62.976	79.348	67.854	92.482
cluster_2	77.851	92.766	98.461	63.578	79.305	53.435
cluster_3	88.634	93.114	87.78	108.772	100.659	83.146

Figura 6: Centroidi

è possibile dare la seguente interpretazione euristica dei cluster:

- cluster_0 : composto da Pokémon con caratteristiche mediamente più deboli
- cluster_1 : composto dai Pokémon mediamente più veloci
- cluster_2 : composto da Pokémon con Attacco e Difesa elevati, ovvero quelle caratteristiche definite come "fisiche" in fase di esplorazione

- cluster_3 : composto da Pokémon con Attacco Speciale e Difesa Speciale elevati, ovvero quelle caratteristiche definite come "speciali" in fase di esplorazione. Inoltre i Pokémon di questo cluster spiccano anche per avere HP elevato .

4.3 K-medoids

Il secondo algoritmo implementato è K-medoids. Anche in questo caso è stata utilizzata la metrica euclidea come distanza. Osservando la Figura 7, si è deciso di impostare l'algoritmo con k=3 come numero di cluster ottimale.

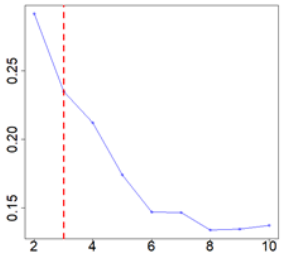

Figura 7: Silhouette K-medoids

Di seguito si riportano i medoidi dei cluster ottenuti:

Row ID	HP	Attack	Defense	Sp_Atk	Sp_Def	Speed
Row177	65	75	70	95	70	95
Row30	90	92	87	75	85	76
Row360	50	50	50	50	50	50

Figura 8: Medoidi

Figura 9: Xatu

Figura 10: Nidoqueen

Figura 11: Snorunt

I medoidi dei cluster ottenuti sono ordinatamente i seguenti: Xatu, Nidoqueen e Snorunt. È quindi possibile dare la seguente interpretazione euristica dei cluster:

- Primo cluster: composto da Pokémon con elevato Attacco Speciale e Velocità
- Secondo cluster: Composto da Pokémon con Hp, Attacco e Difesa più alti della media.
- Terzo cluster: con caratteristiche mediamente più deboli

4.4 Hierarchical-Average

Dopo aver impiegato algoritmi di clustering Prototype-based, si è optato per utilizzare metodi gerarchici di tipo agglomerativo. Si è effettuato un primo tentativo implementando l'algoritmo Hierarchical con metodo average-linkage e distanza euclidea. In Figura 12 viene riportato il dendrogramma.

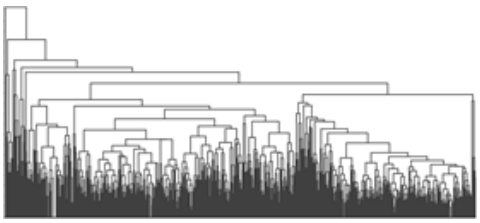

Figura 12: Dendogramma Hierarchical-Average

Nonostante il dendrogramma risulti molto sbilanciato, è stato effettuato un tentativo usando k=3 come numero di cluster di interesse. Di seguito si riportano le medie degli attributi e la dimensione di ogni cluster:

Cluster	HP	Attack	Defense	Sp_At...	Sp_Def	Speed	Size
cluster_0	50	150	50	150	50	150	1
cluster_1	66.455	78.727	139	70.545	143.364	51.818	11
cluster_2	67.765	75.179	69.332	68.769	67.577	66.078	702

Figura 13: Media dei cluster

Come si può osservare le cardinalità dei cluster sono notevolmente differenti. In particolare il cluster 0 è formato da un solo Pokémon: Deoxys.

Figura 14: Deoxys

Deoxys è un Pokémon particolare: non è un outlier, ma presenta quasi tutte le statistiche superiori alla media (tranne *Defense* e *Sp_Def*), il che influenza pesantemente il calcolo della distanza tra esso e gli altri cluster. Il cluster 1, invece, formato da soli 11 elementi, contiene i Pokémon con elevate caratteristiche di difesa, sia normale che speciale.

4.5 Hierarchical-Ward

Dal momento che il risultato ottenuto è insoddisfacente, si è deciso di provare ad utilizzare lo stesso algoritmo di clustering, implementato questa volta con metodo "Ward"e distanza euclidea. Dall'osservazione del dendrogramma, si evince che l'algoritmo Hierarchical-Ward possa essere implementato usando k=3 come numero di cluster di interesse.

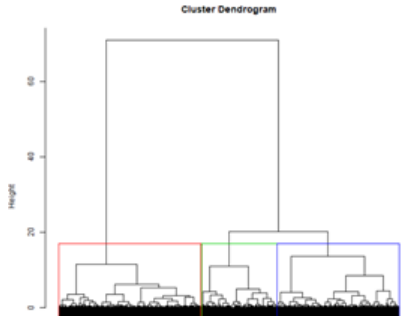

Figura 15: Dendogramma Hierarchical-Ward

Di seguito si riportano le medie degli attributi e la dimensione di ogni cluster:

cluster	HP	Attack	Defense	Sp_At...	Sp_Def	Speed	Size
0	79.238	91.906	79.203	69.105	70.418	72.297	256
1	80.899	89.264	86.616	103.874	97.761	86.126	159
2	50.849	53.749	54.187	50.151	51.823	49.849	299

Figura 16: Media dei Cluster

Utilizzando il metodo Ward le dimensioni dei cluster risultano essere più bilanciate. È possibile dare la seguente interpretazione euristica dei cluster:

- cluster_0 : composto da Pokémon con Attacco fisico elevato e il resto delle caratteristiche nella media
- cluster_1 : composto da Pokémon con Attacco speciale e Difesa speciale più alti della media
- cluster_2 : composto da Pokémon con caratteristiche mediamente più deboli

5 Validazione

Per confrontare i quattro algoritmi di clustering implementati in precedenza sono stati confrontati i valori dell'indice di Silhouette relativi al numero di cluster ritenuto ottimale per ogni algoritmo. Di seguito si riporta la tabella di confronto:

Algoritmo	Clusters	Silhouette
k-means	4	0.2402910
k-medoids	3	0.234894
hierarchical-average	3	0.29821343
hierarchical-ward	3	0.2086679

L'algoritmo caratterizzato dal valore più elevato dell'indice di Silhouette risulta essere Hierarchical-average, che però è stato ritenuto poco soddisfacente. Si decide pertanto di designare l'algoritmo k-means come ottimale. Per quest'ultimo si esegue il test d'ipotesi di Monte Carlo, al fine di stabilire che esista un'effettiva struttura all'interno del dataset e che quindi i gruppi trovati siano significativi e non frutto del caso. Il coefficiente di Silhouette relativo all'algoritmo K-means precedentemente calcolato, viene utilizzato per testare l'ipotesi nulla H_0 (Random Position Hypothesis). Applicando il metodo Monte Carlo, viene generata una distribuzione empirica sulla base di 500 simulazioni sotto l'ipotesi nulla. Con un livello di significatività $\alpha = 0.01$ viene calcolato il quantile della suddetta distribuzione e successivamente confrontato con il coefficiente di Silhouette calcolato in precedenza.

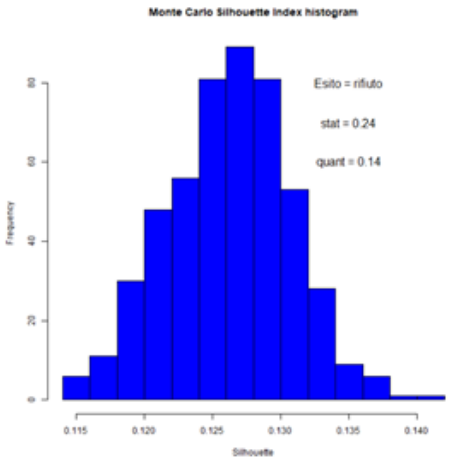

Figura 17: Distribuzione di frequenza per l'indice di Silhouette con Montecarlo

Poiché il quantile ottenuto con Monte Carlo risulta inferiore rispetto a quello della statistica test (ovvero il coefficiente di Silhouette di K-means), si rifiuta l'ipotesi nulla di assenza di struttura e si valida così la soluzione di clustering proposta.

6 PCA

Allo scopo di visualizzare i cluster ottenuti su uno spazio dimensionale ridotto, si è deciso di applicare il metodo delle Principal Component Analysis (PCA). Esso permette di ridurre il numero di dimensioni del dataset di partenza andando a creare nuovi attributi ortogonali come combinazioni lineari di quelli iniziali e cercando di catturare il massimo ammontare di variabilità nei dati. Di seguito si riportano gli autovalori della matrice della covarianza che rappresentano il contributo di ogni variabile nella costruzione delle sei componenti principali.

	PC1	PC2	PC3	PC4	PC5	PC6
HP	0.44	-0.13	0.30	-0.68	0.36	-0.33
Attack	0.43	-0.16	0.63	0.20	-0.22	0.55
Defense	0.37	-0.59	-0.15	0.48	-0.10	-0.50
Sp_Atk	0.44	0.36	-0.29	-0.25	-0.72	-0.11
Sp_Def	0.44	-0.05	-0.61	0.00	0.42	0.50
Speed	0.31	0.69	0.17	0.46	0.34	-0.27

Figura 18: Autovettori relativi agli attributi

Successivamente si riporta una tabella con le deviazioni standard, la porzione di varianza spiegata da ciascuna componente e la varianza cumulata.

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6
Standard deviation     1.6700 1.0379 0.9120 0.7895 0.6364 0.52325
Proportion of Variance 0.4648 0.1795 0.1386 0.1039 0.0675 0.04563
Cumulative Proportion  0.4648 0.6443 0.7830 0.8869 0.9544 1.00000
```

Figura 19: Statistiche delle PCA

Al fine di poter rappresentare graficamente i dati in uno spazio bidimensionale vengono selezionate le prime due componenti principali, che insieme spiegano circa il 64% della varianza dei dati. Aggiungendo una terza dimensione si potrebbe raggiungere circa il 78% della porzione di varianza spiegata. Le due nuove variabili rappresentate nel grafico sono quindi:

- PCA1: componente principale caratterizzata unicamente da coefficienti positivi, in cui tutti gli attributi di partenza contribuiscono approssimativamente nella stessa misura
- PCA2: componente principale in cui gli attributi che contribuiscono in maggior misura sono '*Defense*' (negativamente) e '*Speed*' (positivamente).

In Figura 20 sono rappresentati i cluster trovati con l'algoritmo K-means implementato precedentemente.

Figura 20: Scatter plot PCA1 e PCA2

7 Clustering caratteristiche fisiologiche

7.1 Preprocessing

Per rispondere alla seconda domanda di ricerca sono state selezionate le variabili riguardanti le principali caratteristiche fisiologiche dei Pokémon:

- *Type_1*: descrive la natura del Pokemon
- *Height_M*: altezza del Pokémon espressa in metri
- *Weight_KG*: peso del Pokémon espresso in Kg
- *Body_Style*: descrive la forma del Pokèmon (a quattro zampe, insetto, con la coda o senza, ecc.)

A differenza del problema affrontato nella prima sezione, qui le variabili sono di tipo misto, sia categoriche (*Type_1*,*Body_Style*) che numeriche (*Height_M*, *Weight_Kg*). Per questo motivo è

necessario introdurre un nuovo tipo di misura di distanza: la distanza di Gower. La distanza di Gower tra due record i e j è definita come:

$$d(i,j) = \frac{\sum_k \delta_{ijk} d_{ijk}}{\sum_k \delta_{ijk}}$$

Ove k è il numero di attributi del dataset e δ_{ijk} vale:

- $\delta_{ijk} = 0$ se $x_{ik} = NA$ o $x_{jk} = NA$
- $\delta_{ijk} = 0$ se la variabile è asimmetrica binaria e $x_{ik} = x_{jk} = 0$ o $x_{ik} = x_{jk} = False$
- $\delta_{ijk} = 1$ negli altri casi

Per quanto riguarda d_{ijk},cioè la distanza tra i record i e j per l'attributo k è caratterizzato dalla prima formula se l'attributo è numerico e dalla seconda se è categorico.

$$d_{ijk} = \frac{|x_{ik} - x_{jk}|}{R_k}$$

$$d_{ijk} = \begin{cases} 0 & \text{se} \quad x_{ik} = x_{jk} \\ 1 & \text{altrimenti} \end{cases}$$

Dove R_k è il range dell'attributo k. Per come è definita la distanza di Gower, non è necessario normalizzare, dal momento che tutte le distanze calcolate varieranno in un intervallo definito da 0 a 1. Per migliorare la comprensione del dataset e facilitarne l'interpretabilità, si è deciso di ridurre i livelli della variabile categorica*Body_ Style*. Da 14 livelli iniziali si è arrivati per aggregazione a 5 livelli finali, ovvero: bipedal, head, wings, quadruped, other. Lo stesso procedimento non è stato adoperato per la variabile *Type_ 1*, in quanto non è stato possibile trovare una logica di aggregazione. Infine, sono stati eliminati gli outliers delle variabili *Height_ M* e *Weight_ Kg*, ovvero i record non appartenenti all'intervallo $(Q1_i - 1.5 * IQR_i, Q3_i + 1.5 * IQR_i)$. In questo modo si riduce notevolmente il range di variazione delle due variabili numeriche.

7.2 K-medoids

Come algoritmo di clustering si è deciso di utilizzare K-medoids dal momento che questo algoritmo funziona in modo ottimale anche con distanze diverse da quelle di tipo Minkowski. Per determinare il numero ottimale di cluster è stato studiato l'andamento dell'indice di Silhouette.

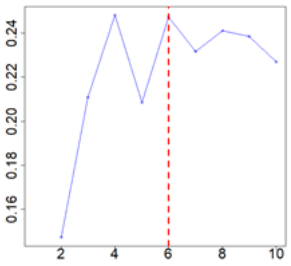

Figura 21: Silhouette K-medoids

Dal grafico si può osservare che per k=4 e k=6 si hanno i valori di Silhouette più elevati, pressoché identici. Per questo motivo si decide di utilizzare k=6 come numero di cluster ottimale. Di seguito si riportano i medoidi ottenuti:

Row ID	Type_1	Heigh...	Weig...	group...	size
Row510	Grass	0.61	10.5	bipedal	132
Row268	Bug	1.19	31.6	wings	67
Row607	Ghost	0.61	13	head	92
Row317	Water	0.79	20.8	other	100
Row286	Normal	0.79	24	quadruped	132
Row61	Water	1.3	54	bipedal	121

Figura 22: Medoidi

Figura 23: Pansage

Figura 24: Poliwrath

Figura 25: Carvanha

Figura 26: Slakoth

Figura 27: Lampent

Figura 28: Dustox

è possibile fare le seguenti osservazioni sui cluster:

- Sono presenti due cluster di bipedali dove il primo è composto da Pokémon aventi altezza e peso più bassi rispetto al secondo.
- Sono presenti due cluster rappresentati dal tipo acqua, uno caratterizzato da una fisionomia che permette al pokémon di stare fuori dall'acqua, mentre l'altro caratterizzato da fisionomie più particolari.

8 Conclusioni

Per quanto riguarda la prima domanda di ricerca, incentrata sulle caratteristiche di battaglia dei Pokémon, si può concludere che i Pokémon tendono a non seguire il fenomeno del glass cannon, in quanto non presentano particolari asimmetrie tra le caratteristiche combattive. Il metodo di clustering che è risultato migliore, considerati l'indice di Silhouette e l'interpretabilità dei risultati, è K-means. Questo algoritmo suddivide i Pokémon in quattro gruppi principali: pokémon deboli, pokemon veloci, pokemon con caratteristiche fisiche elevate e pokemon con caratteristiche speciali più forti. Rispetto alla seconda domanda di ricerca, riguardante le caratteristiche fisiologiche dei Pokémon, la cluster analysis non porta ad un risultato soddisfacente e di facile interpretazione. Questo è dovuto al fatto che i Pokèmon presentano strutture e dimensioni molto eterogenee, essendo un prodotto di immaginazione atto a stimolare la curiosità dei fruitori.

Riferimenti

[1] Bulbapedia. URL: https://bulbapedia.bulbagarden.net/wiki/Main_Page.

[2] Dataset. URL: https://www.kaggle.com/alopez247/pokemon.

[3] Knime. URL: https://www.knime.com/.

[4] Gower Measure. URL: https://www.rdocumentation.org/packages/StatMatch/versions/1.2.0/topics/gower.dist.

[5] clValid Package. URL: http://cran.us.r-project.org/web/packages/clValid/vignettes/clValid.pdf.

[6] PCA. URL: https://www.agnesevardanega.eu/wiki/r/analisi_esplorativa/analisi_in_componenti_principali.

[7] RStudio. URL: https://www.rstudio.com.

[8] TassoDiCattura. URL: https://wiki.pokemoncentral.it/Tasso_di_cattura#Probabilit.C3.A0_approssimata.