

Analisi dei Pokemon e loro battaglie

Caliaro Lorenzo 832652

Leanza Alfio 910235

Tavecchio Riccardo 852211

Abstract:

In questo studio, ci siamo posti come obiettivo di identificare le caratteristiche che influenzano significativamente le vittorie nei combattimenti Pokémon e di comprendere gli attributi distintivi che rendono un Pokémon leggendario. Per raggiungere questi obiettivi, abbiamo condotto analisi approfondite utilizzando tre dataset distinti, ognuno fornendo informazioni sulle statistiche, le abilità e gli attributi dei Pokémon. Attraverso un'approfondita esplorazione e l'applicazione di tecniche statistiche, siamo stati in grado di individuare i fattori chiave che incidono sul successo in battaglia. Inoltre, abbiamo analizzato dettagliatamente le caratteristiche che contraddistinguono i Pokémon leggendarie, contribuendo a una migliore comprensione di questa categoria particolare. I risultati ottenuti forniscono insights preziosi per gli appassionati di Pokémon e per chiunque sia interessato a comprendere più a fondo le dinamiche dei combattimenti Pokémon.

Data acquisition:

In questa prima fase del progetto, si è acquisito il dataset dei Pokémon tramite API dall'endpoint "https://pokeapi.co/". Il dataset ottenuto è stato salvato in una collezione MongoDB.

Le variabili osservate sono le seguenti:

- **Name:** nome del Pokémon;
 - **Generation:** generazione di appartenenza;
 - **Height:** altezza in cm;
 - **Weight:** peso in kg;
 - **Types:** tipo a cui appartiene un Pokémon (può essere monotipo, o doppio tipo);
 - **HP:** punti vita o salute, definisce quanto danno può sopportare un Pokémon prima di andare KO;
 - **Attack:** attacco base;
 - **Defense:** difesa base (resistenza al danno contro attacchi base);
 - **Special Attack:** attacco speciale;
 - **Special Defense:** resistenza al danno da attacchi speciali;
 - **Speed:** velocità associata a ciascun Pokémon, determina chi attacca per primo in una lotta e l'abilità di poter schivare un attacco;
 - **Abilities:** serie di abilità possedute;
 - **Learnable moves:** mosse apprendibili;
 - **Hidden ability:** abilità nascosta.

Come si può intuire dal nome delle colonne, il dataset mostra tutti i Pokemon con le loro caratteristiche e statistiche.

In particolar modo, le colonne HP, Attack, Defense, Special Attack, Special Defense, Speed sono in formato numerico, e mostrano valori che, una volta aggregati, mostrano il totale delle abilità di ciascun Pokémon; di conseguenza è stata aggiunta la colonna "Total Stats". Inoltre sono state immesse ulteriori colonne che conteggiano il numero di 'Abilities', 'Learnable Moves'. Infine il dataset è stato ripulito dalle righe che presentavano valori nulli nella colonna "Generation"; sono stati riscontrati valori nulli anche nella colonna "Hidden Ability", ma in questo caso rappresentano la mancanza di abilità nascoste.

Successivamente sono stati oggetto di studio altri due dataset in formato CSV: "Pokemon 2" "combats".

Il dataset combats mostra le seguenti colonne:

- **First_pokemon**: indice del primo Pokémon lottatore;
- **Second_pokemon**: indice del secondo Pokémon in gara;
- **Winner**: indice del vincitore.

Esso è composto da 50000 righe e ognuna di esse corrisponde ad una lotta fra due Pokemon. Sono state svolte operazioni di pre-processing come la rimozione di 1952 righe duplicate, che rappresentavano la stessa battaglia con lo stesso esito. Inoltre sono state rimosse righe con valori mancanti e valori nulli

Il dataset "Pokemon 2" presenta le seguenti colonne: **#, Name, Type 1, Type 2, HP, Attack, Defense, Sp. Atk, Sp. Def, Speed, Generation, Legendary**. Proprio come il primo dataset "Pokemon", esso mostra le caratteristiche di ogni Pokémon con anche l'aggiunta della colonna Legendary, che specifica tramite un valore booleano se il Pokémon è leggendario o meno. Inoltre è stato riscontrato un valore nullo sulla colonna "Name" ed è stato eliminato.

Esso serve a identificare i Pokémon presenti nel dataset "combats". In questo dataset, per ulteriori analisi successive, sono state inserite la colonna di "**partite giocate**", "**partite vinte**" e "**% vittorie**" su ogni Pokemon, calcolate tramite il dataset "combats" in corrispondenza degli indici. Anche questi due dataset sono stati salvati in collezioni di MongoDB.

STORAGE

I dataset ottenuti sono stati memorizzati come collezioni nel database MongoDB, consentendo un accesso rapido e agevole ogni volta che si desidera utilizzarli nel codice. Questa strategia di archiviazione agevola la gestione e il recupero affidabile dei dati attraverso il tempo, garantendo una maggiore efficienza nello sviluppo e nell'esecuzione del codice.

DATA INTEGRATION/ENRICHMENT

A questo punto si passa alla fase di integrazione dei dati. Come da teoria, ci sono diversi modi per integrare i dati da diverse fonti, e uno di questi è il "Data Enrichment". Si sono uniti i due dataset "Pokémon" e "Pokémon 2" tenendo solo le colonne in comune, e quelle di interesse per l'analisi. Si punta quindi alla qualità dei dati, rimuovendo colonne inutili e mantenendo solo le colonne rilevanti,

creando un dataset consolidato che può essere utilizzato per eseguire analisi più approfondite, identificare correlazioni e ottenere una comprensione più completa del panorama dei Pokémon. Il dataset risultante presentava 14 valori nulli nella colonna “% Winner” sostituiti da dati predetti tramite un algoritmo di machine learning, dividendo in train e test data.

Si nota che nel dataset “Pokemon 2”, sono presenti anche le evoluzioni dei vari Pokémon, le quali non sono state prese in considerazione nel dataset finale, poiché non presenti nel primo dataset. Inoltre si nota che in “Pokémon 2”, in 25 Pokémon ci sono dei bias nei valori di alcune caratteristiche. Dopo aver controllato sul sito ufficiale dei Pokémon, si nota che i valori corretti sono quelli appartenenti al dataset “Pokemon” e quindi si è provveduto a sostituire i dati non accurati.

DATA EXPLORATION

Per quanto riguarda il dataset finale, si hanno i seguenti missing values:

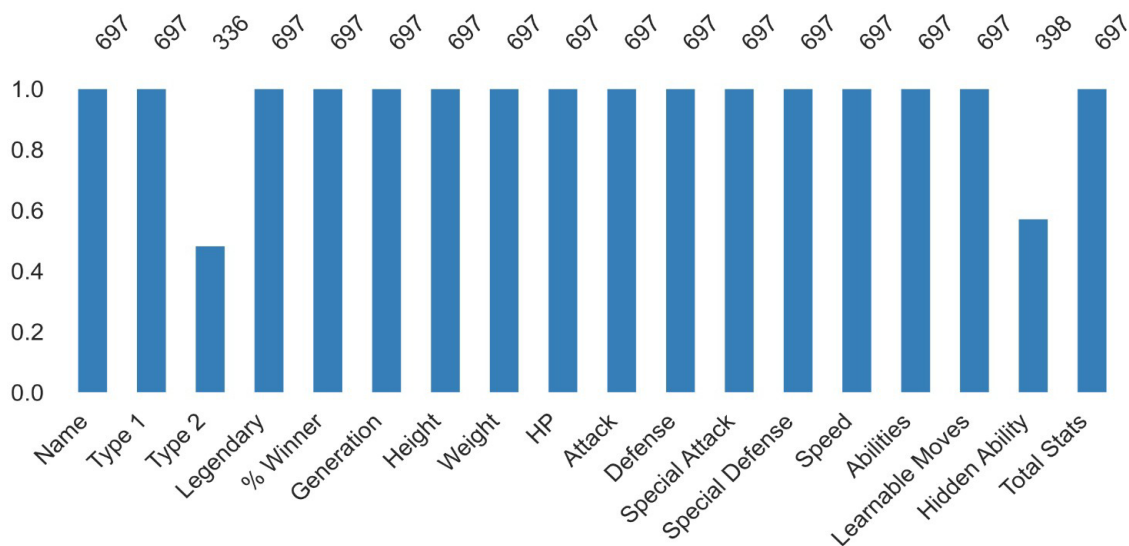


Immagine 1 : missing values dataset finale

La presenza di valori mancanti nelle colonne “Type 2” e “Hidden Ability” non intende un dataset incompleto o una mancanza di imputazione, poiché il valore nullo identifica l’assenza di un secondo tipo o la mancanza di una abilità nascosta.

Matrice di correlazione:

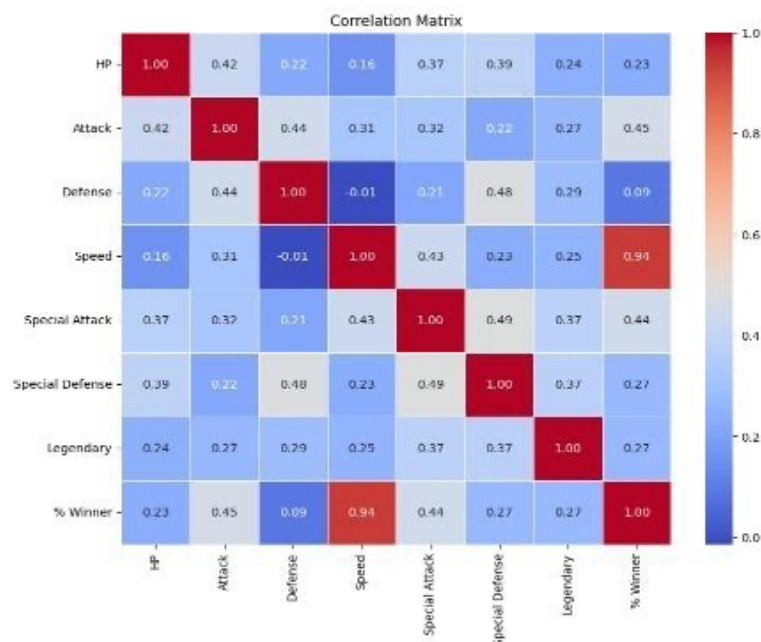


Immagine 2: Matrice di correlazione tra le variabili del dataset finale

Come si può notare, c'è una forte correlazione positiva (pari a 0.94) tra la velocità ("Speed") di un Pokémon e la variabile "% Winner". Ciò significa che un Pokémon con una velocità alta vincerà quasi sempre le battaglie. Esso può essere spiegato dal fatto che, durante una lotta, il Pokémon con una maggiore velocità avrà la possibilità di attaccare per primo, controllare al meglio il campo di lotta e infine aumentare la probabilità di schivare un attacco non subendo alcun danno.

La seguente immagine mostra quanto appena spiegato, ovvero le percentuali di vittoria dei Pokémon con la loro velocità. Si nota che i Pokémon con un'alta percentuale di vittorie hanno anche una velocità molto alta rispetto ai Pokémon con una bassa percentuale di vittorie.

| | Name | % Winner | Speed |
|-----|------------|----------|-------|
| 452 | weavile | 97.41 | 125 |
| 136 | aerodactyl | 96.32 | 130 |
| 640 | greninja | 95.76 | 122 |
| 209 | sneasel | 94.34 | 115 |
| 481 | darkrai | 93.64 | 125 |
| ... | ... | ... | ... |
| 212 | slugma | 3.39 | 20 |
| 564 | solosis | 3.23 | 20 |
| 169 | togepi | 2.61 | 20 |
| 260 | silcoon | 2.38 | 15 |
| 207 | shuckle | 0.00 | 5 |

Immagine 3: percentuali di vittorie per ogni Pokémon. La velocità è un fattore incisivo

Inoltre andando a valutare le altre variabili notiamo che i seguenti valori hanno una discreta relazione positiva:

- ATTACCO vs. DIFESA;
- ATTACCO SPECIALE vs. DIFESA SPECIALE;
- DIFESA vs. DIFESA SPECIALE.

Questo, intuitivamente, ci dice che un Pokémon che ha un attributo di attacco molto alto, probabilmente, avrà un valore di difesa alto (stessa considerazione vale anche per le altre due coppie di variabili).

Questa correlazione è logica se teniamo conto dell'equilibrio comune nella progettazione dei Pokémon. Un Pokémon con un attacco aggressivo e potente può anche avere una buona resistenza e difesa, che gli permettono di resistere ai contrattacchi del suo avversario. Allo stesso modo, i Pokémon specializzati in attacchi speciali possono beneficiare di difese speciali più forti per impedire tali mosse da parte dei nemici. Questo equilibrio aiuta a garantire che ogni Pokémon abbia una combinazione di attributi che lo rendono efficace in diverse situazioni di combattimento.

Vale la pena notare, tuttavia, che non tutte le statistiche dei Pokémon sono strettamente correlate. Alcuni Pokémon potrebbero specializzarsi in un certo tipo di attacco o difesa, sacrificando altre statistiche per migliorare le proprie prestazioni in quella specifica area.

Si valutano ora gli **scatter plot** dei vari Pokemon divisi per leggendari e non leggendari:

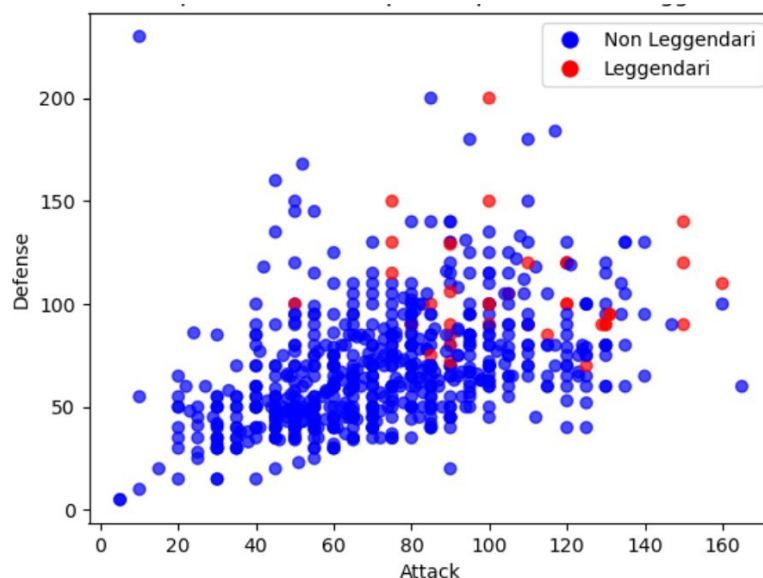


Immagine 4: Scatter plot tra attacco e difesa per Pokémon leggendari e non leggendari

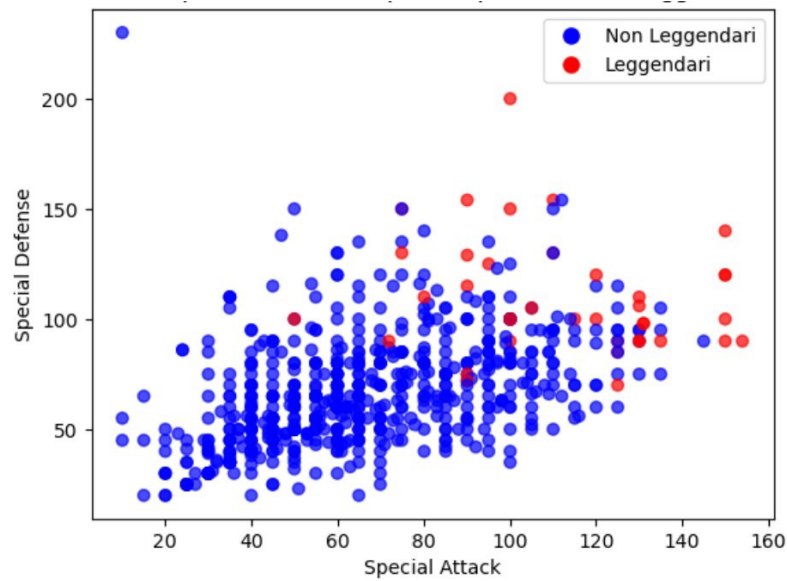


Immagine 5: Scatter plot tra attacco speciale e difesa speciale per Pokémon leggendari e non leggendari

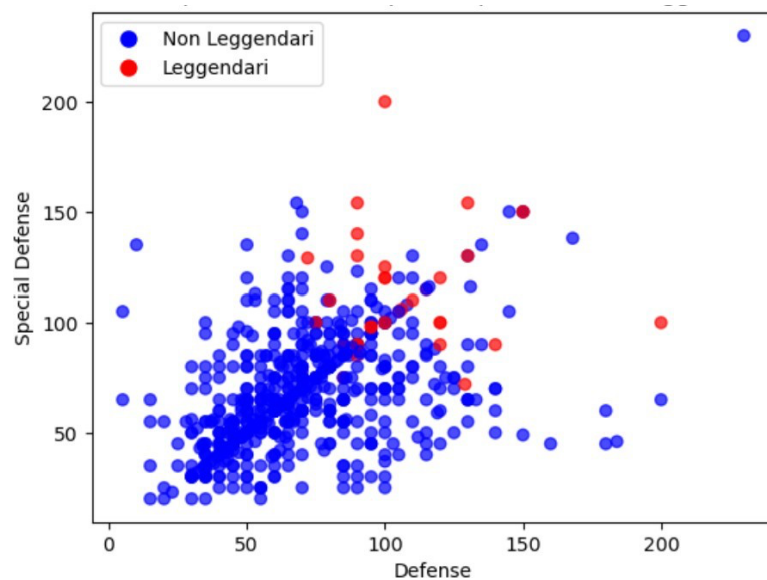


Immagine 6: Scatter plot tra difesa e difesa speciale per Pokémon leggendari e non leggendari

Come si può notare dagli scatterplot, tipicamente i valori associati ai diversi attributi dei Pokémon leggendari sono più alti, in media, rispetto a quelli normali.

Tuttavia, è altrettanto evidente che non si può discriminare la natura leggendaria di un Pokémon visivamente.

Si va quindi a rappresentare i **boxplot** per ogni caratteristica, divisa per legendary e non legendary:

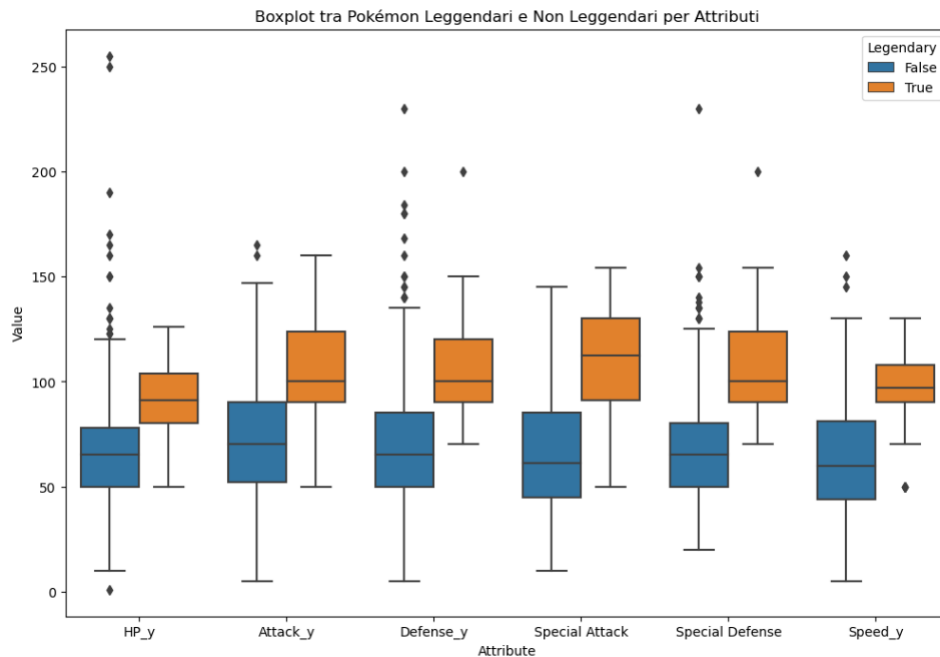


Immagine 6: Box plot tra Pokémon leggendari (Arancio) e non leggendari(blu) per attributi.

Come supposto prima, ora in modo evidente osservando i boxplot i Pokémon leggendari sono progettati per essere molto potenti e avere statistiche più alte rispetto alla maggior parte dei Pokémon non leggendari. Questa differenza è una parte importante della loro identità nel mondo dei Pokémon, riflettendo il loro status unico e spesso mitico. D'altra parte, è interessante notare che molti Pokémon normali possono mostrare valori anomali in determinati attributi rispetto ai loro compagni della stessa categoria.

Nel complesso, ciò suggerisce che i tratti dei Pokémon leggendari sono più omogenei rispetto a quelli dei Pokémon normali, mentre i valori degli attributi dei Pokémon normali possono mostrare una maggiore variabilità e diversità. Questa differenza probabilmente riflette la loro diversa natura e ruolo nel mondo dei Pokémon, con i Pokémon leggendari che incarnano la perfezione e la potenza, mentre i Pokémon normali rappresentano un'ampia varietà di forme, dimensioni e abilità.

DATA QUALITY

I dati che sono stati analizzati provenienti dai 3 dataset, una volta puliti, risultano essere completi e coprono tutte le informazioni necessarie. Inoltre sono consistenti in quanto andando a comparare i valori con quelli reali dai vari siti ufficiali dei Pokémon, risultano essere corretti. Dopo aver eliminato i vari duplicati possiamo affermare l'unicità dei dati. I dati dopo le operazioni di arricchimento del dataset risultano essere rilevanti in quanto le colonne non di interesse sono state scartate.

SVILUPPI FUTURI

Il presente dataset si focalizza attentamente sui diversi Pokémon prima delle loro mega-evoluzioni, analizzando dettagliatamente le loro caratteristiche. Un possibile sviluppo futuro potrebbe estendere la ricerca anche alle loro forme mega-evolutive, esplorando i cambiamenti nelle loro prestazioni.

In aggiunta, si potrebbe considerare l'implementazione di modelli predittivi per stimare la percentuale di vittoria di Pokémon non inclusi nel dataset dei combattimenti. Questa prospettiva consentirebbe di ampliare ulteriormente la comprensione delle capacità e delle potenzialità di ciascun Pokémon, contribuendo a una visione più completa e predittiva del loro rendimento in situazioni di combattimento.

RUOLI DEI PARTECIPANTI:

- **Caliaro Lorenzo:** identificazione e acquisizione delle fonti di dati necessarie per il progetto;
- **Leanza Alfio:** responsabile della progettazione e dell'implementazione dell'infrastruttura di archiviazione dei dati;
- **Tavecchio Riccardo:** garanzia della qualità dei dati durante tutto il processo e unificazione dei dati.

Durante il processo esplorativo tutti i partecipanti hanno contribuito dando una propria prospettiva tramite dibattiti e confronti.