

Predicting Pokémon Fights Outcomes

Marco Belotti (793675)

Francesco Bombarda (794976)

Antonio Vivace (793509)

A.A. 2017-2018





Dominio Applicativo

- L'ambito entro cui si colloca il progetto riguarda una delle feature più importanti dei videogiochi della serie di successo Pokémon, ovvero i combattimenti.
- Queste Battles si basano principalmente su alcune statistiche tra cui:
 - **PS(HP)**. Indica l'energia vitale di un Pokémon.
 - **Attacco**. Statistica da cui dipende l'entità dei danni che il Pokémon può provocare con attacchi fisici.
 - **Difesa**. Indica la resistenza di un Pokémon agli attacchi fisici.
 - **Attacco Speciale**. Statistica da cui dipende l'entità dei danni che il Pokémon può provocare con attacchi speciali.
 - **Difesa Speciale**. Resistenza di un Pokémon agli attacchi speciali.
 - **Velocità**. Indica la rapidità di un Pokémon.

Obiettivo del progetto

- Il progetto in esame si propone dunque d'individuare un modello di Machine Learning in grado di classificare correttamente il risultato di uno scontro tra Pokémon, il tutto garantendo livelli di accuratezza e performance il più elevati possibili, ricavando le informazioni necessarie per la scelta da vari dataset disponibili in rete, i quali saranno descritti nelle slide successive



Raccolta dei dati

- Dopo aver scelto il dominio applicativo di riferimento e gli obiettivi principali dello stesso, si è reso necessario individuare in rete alcuni dataset contenenti informazioni utili a soddisfare l'analisi richiesta. Queste informazioni sono state ricavate dalla piattaforma Kaggle, la quale ci ha messo a disposizione i seguenti dataset:
 - **pokemon.csv**
 - **combats.csv**
 - **pokemonTypeComp.csv**
 - **test.csv.**

kaggle

Dimensioni di qualità singoli dataset

- Al fine di effettuare una prima analisi sui dati raccolti, si è deciso di andare a monitorare quali fossero le misure di qualità relative a completezza ed unicità sui vari dataset da integrare.

- Ecco alcuni degli aspetti più rilevanti:

- Completezza pokemon.csv

id	Name	Type.1	Type.2	HP	Attack	Defense	Sp.Atk	Sp.Def	Speed	Generation	Legendary
0.00000	0.00125	0.00000	0.48250	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

- Unicità pokemon.csv

id	Name	Type.1	Type.2	HP	Attack	Defense	Sp.Atk	Sp.Def	Speed	Generation	Legendary
1.00000	1.00000	0.02250	0.02375	0.11750	0.13875	0.12875	0.13125	0.11500	0.13500	0.00750	0.00250

- Completezza combats.csv

First_pokemon	Second_pokemon	winner
0	0	0

- Unicità tra coppie in combats.csv

```
> # Unicità percentuale tra le coppie di combattenti
> unique_versus <- sum(length(unique(combats$First_pokemon, combats$Second_pokemon)))
> Uniqueness_versus <- unique_versus/length(combats$First_pokemon)
> Uniqueness_versus
[1] 1
```

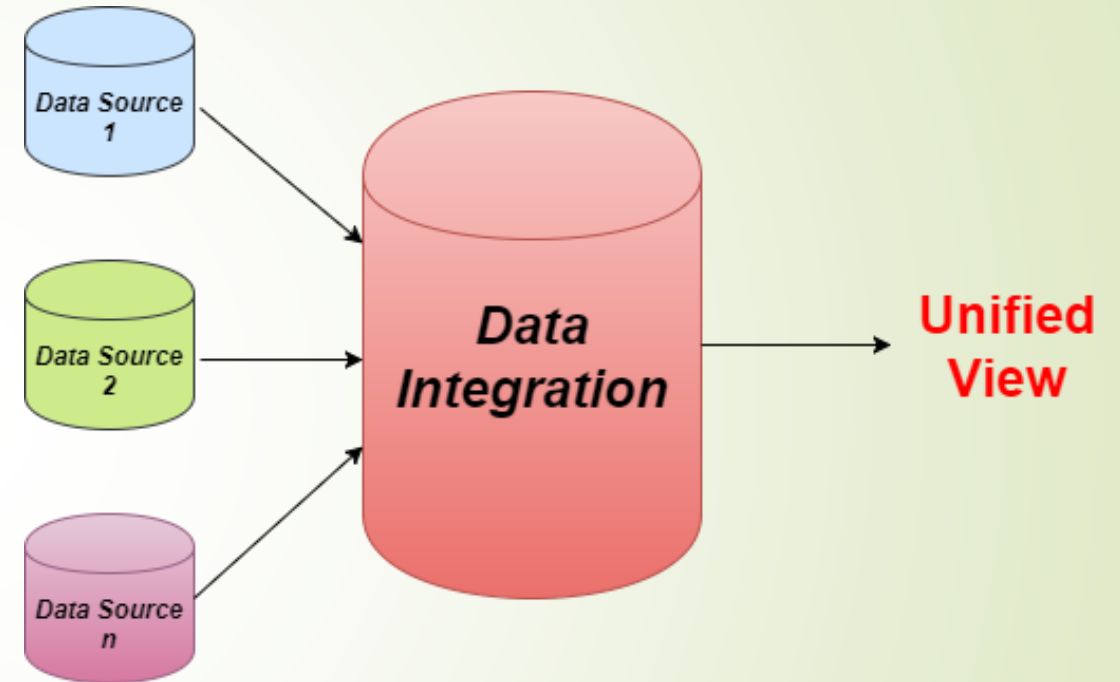

Aspetti evidenziati dalle misure di qualità

- In pokemon.csv vi è la mancanza del campo Name associato al Pokémon con id 63
- L'attributo Type. 2 in pokemon.csv risulta mancante nel 48% dei casi
- Id e Name in pokemon.csv sono effettivamente univoci
- In combats.csv non vi sono valori mancanti
- Ogni coppia di Pokémon in combats.csv compare una ed una sola volta



Dataset coinvolti nell'integrazione

- L'integrazione ha riguardato i file **pokemon.csv**, **combats.csv** e **pokemonTypeComp.csv**. Attraverso la loro integrazione è stato possibile mettere in relazione all'interno di una stessa collezione tutti i parametri caratteristici dei Pokémon contendenti, oltre alle informazioni riguardo la tipologia e l'esito del combattimento.
- L'obiettivo è infatti quello d'individuare la relazione esistente tra le vittorie nei combattimenti, le statistiche dei vari Pokémon ed il tipo del Pokémon stesso.





Integrazione

- L'integrazione per i dataset pokemon.csv e combats.csv ha coinvolto tutti gli attributi disponibili ad eccezione dell'attributo secondo tipo contenuto in pokemon.csv.
- Sono inoltre stati calcolati alcuni parametri aggiuntivi, utili a sottolineare delle differenze tra le caratteristiche peculiari dei Pokémon tra cui attacco, difesa e velocità. È stata inoltre prevista l'aggiunta dell'attributo **winner_first_label** utile per sapere se il vincitore è il primo dei due Pokémon (dove il primo Pokémon è anche colui che sferra il primo attacco).

Integrazione

- L'integrazione con il dataset pokemonTypeComb.csv ha portato ad aggiungere al file integrato un ulteriore parametro, utile per le analisi successive, cioè l'attributo advantage, che possiamo ricavare dal dataset "pokemonTypeComp.csv" considerando le tipologie dei due Pokémon contendenti.
- Terminato il procedimento d'integrazione, la tabella contenente tutte le informazioni è stata esportata in formato csv e nominata "integrated".

Defender \ Attacker	Normal	Fire	Water	Grass	Electric	Ice	Fighting	Poison	Ground	Flying	Psychic	Bug	Rock	Ghost	Dragon	Dark	Steel	Fairy
Normal	1	1/2	1/2	2	2	1	1	1	1	1	1	1	1	0	1	1	1	1
Fire	1/2	1	2	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1
Water	2	1/2	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
Grass	1/2	2	1/2	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1
Electric	1	1	2	1/2	1	1	1	0	2	1	1	1	1	1	1	1	1	1
Ice	1/2	1/2	2	1	1	1	1	2	2	1	1	1	1	1	2	1	1	1
Fighting	2	1	1	1	1	2	1	1/2	1/2	1/2	1/2	2	0	1	2	2	1/2	1
Poison	1	1	1	2	1	1	1	1	1/2	1/2	1	1	1	1	0	2	1	1
Ground	1	2	1	1/2	2	1	1	1	1	0	1	1	2	1	1	1	2	1
Flying	1	1	1	2	1/2	1	1	1	1	1	1	2	1/2	1	1	1	1	1
Psychic	1	1	1	1	1	1	2	2	1	1	1	1	1	1	0	1	1	1
Bug	1/2	1	2	1	1	1	1/2	1/2	1/2	1	2	1	1	1	1	2	1/2	1/2
Rock	1	2	1	1	1	2	1/2	1/2	1	2	1	1	1	1	1	1	1	1
Ghost	0	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1
Dragon	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	0
Dark	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1
Steel	1/2	1/2	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	2
Fairy	1/2	1	1	1	1	2	1	1	1	1	1	1	1	2	2	1	1	1


Dimensioni di qualità dataset integrato

- Sono state valutate alcune misure di qualità, anche in questo caso in riferimento a completezza ed unicità sui singoli attributi del dataset integrato.
- Completezza per il dataset integrato

First_pokemon	Second_pokemon	Winner	First_pokemon_name	Second_pokemon_name
0.00000	0.00000	0.00000	0.00112	0.00104
First_pokemon_attack	Second_pokemon_attack	Diff_attack	First_pokemon_defense	Second_pokemon_defense
0.00000	0.00000	0.00000	0.00000	0.00000
Diff_defense	First_pokemon_sp_defense	Second_pokemon_sp_defense	Diff_sp_defense	First_pokemon_sp_attack
0.00000	0.00000	0.00000	0.00000	0.00000
Second_pokemon_sp_attack	Diff_sp_attack	First_pokemon_speed	Second_pokemon_speed	Diff_speed
0.00000	0.00000	0.00000	0.00000	0.00000
First_pokemon_HP	Second_pokemon_HP	Diff_HP	First_pokemon_type	Second_pokemon_type
0.00000	0.00000	0.00000	0.00000	0.00000
First_pokemon_legendary	Second_pokemon_legendary	winner_first_label	advantage	
0.00000	0.00000	0.00000	0.00000	

- Unicità per il dataset integrato

First_pokemon	Second_pokemon	Winner	First_pokemon_name	Second_pokemon_name
0.01568	0.01568	0.01566	0.01568	0.01568
First_pokemon_attack	Second_pokemon_attack	Diff_attack	First_pokemon_defense	Second_pokemon_defense
0.00222	0.00222	0.00598	0.00206	0.00206
Diff_defense	First_pokemon_sp_defense	Second_pokemon_sp_defense	Diff_sp_defense	First_pokemon_sp_attack
0.00734	0.00184	0.00184	0.00628	0.00208
Second_pokemon_sp_attack	Diff_sp_attack	First_pokemon_speed	Second_pokemon_speed	Diff_speed
0.00208	0.00634	0.00216	0.00216	0.00550
First_pokemon_HP	Second_pokemon_HP	Diff_HP	First_pokemon_type	Second_pokemon_type
0.00184	0.00184	0.00658	0.00036	0.00036
First_pokemon_legendary	Second_pokemon_legendary	winner_first_label	advantage	
0.00004	0.00004	0.00004	0.00008	



Scelta del modello di Machine Learning

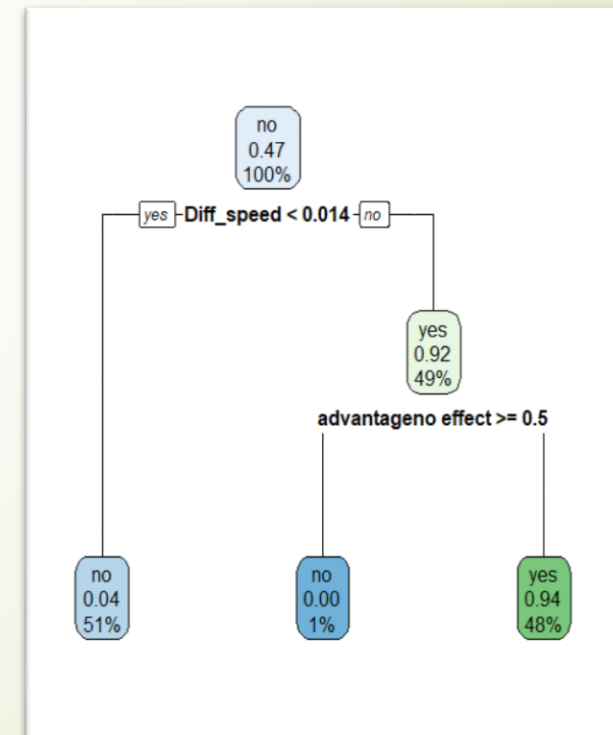
- Il modello scelto per risolvere il problema di classificazione presentato è basato su di un albero di decisione, i motivi che ci hanno spinti ad adottare tale metodologia sono :
 - **Semplicità:** Indubbiamente gli alberi di decisione sono facili da capire e da eseguire
 - **Controllo:** L'uomo può facilmente verificare come la macchina giunge alla decisione ed eventualmente dissentire
 - **Problematica in esame:** Gli alberi decisionali sono notoriamente poco adatti a modellare problemi complessi, tuttavia questo non riguarda la problematica presentata

Modello di Machine Learning

Dopo aver suddiviso il dataset integrato, nelle porzioni di training e testing

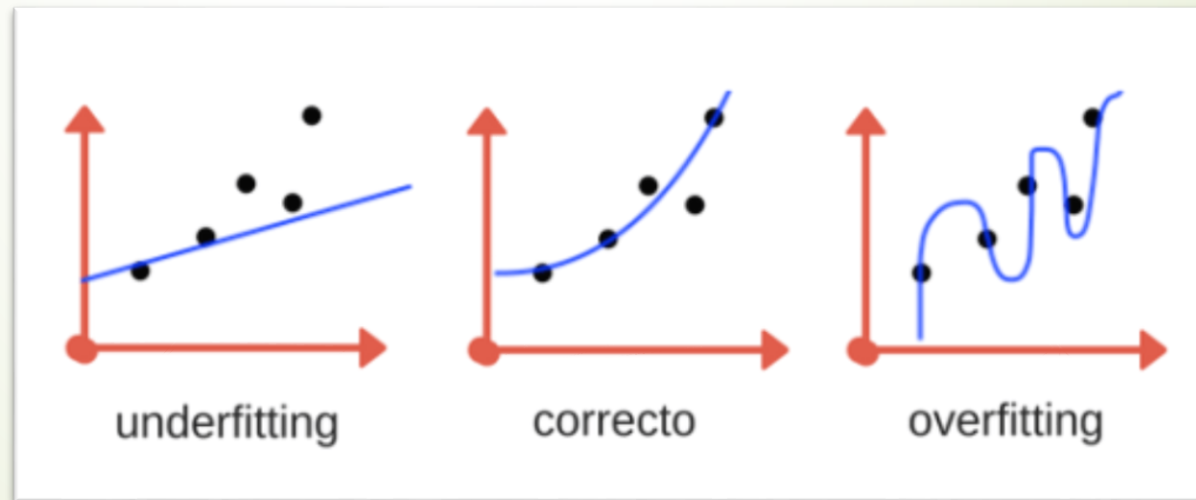
È stato creato il modello (a partire dai dati contenuti nella porzione di training), limitatamente ad alcuni attributi, riportati di seguito:

- winner_first_label
- Diff_attack
- Diff_defense
- Diff_sp_defense
- Diff_sp_attack
- Diff_speed
- Diff_HP
- First_pokemon_legendary
- Second_pokemon_legendary
- Advange (In seguito rimosso)

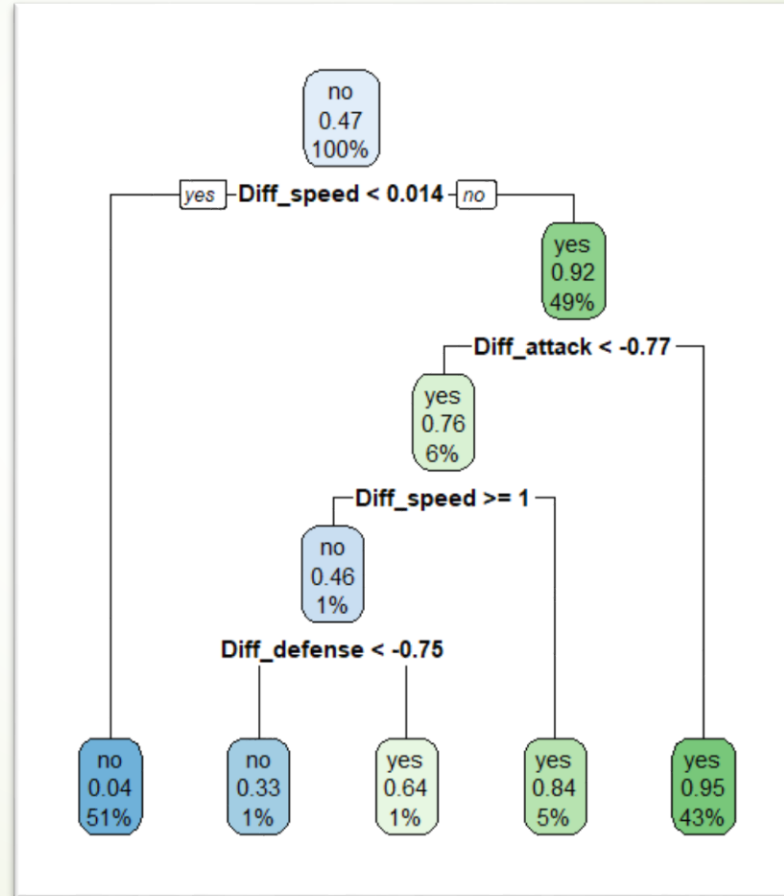


Problematiche riscontrate

- Problema: possibile overfitting, visto il numero limitato di parametri utilizzati per prendere la decisione e l'elevata qualità delle tipiche misure di performance
- Soluzione: Rimozione del parametro advantage

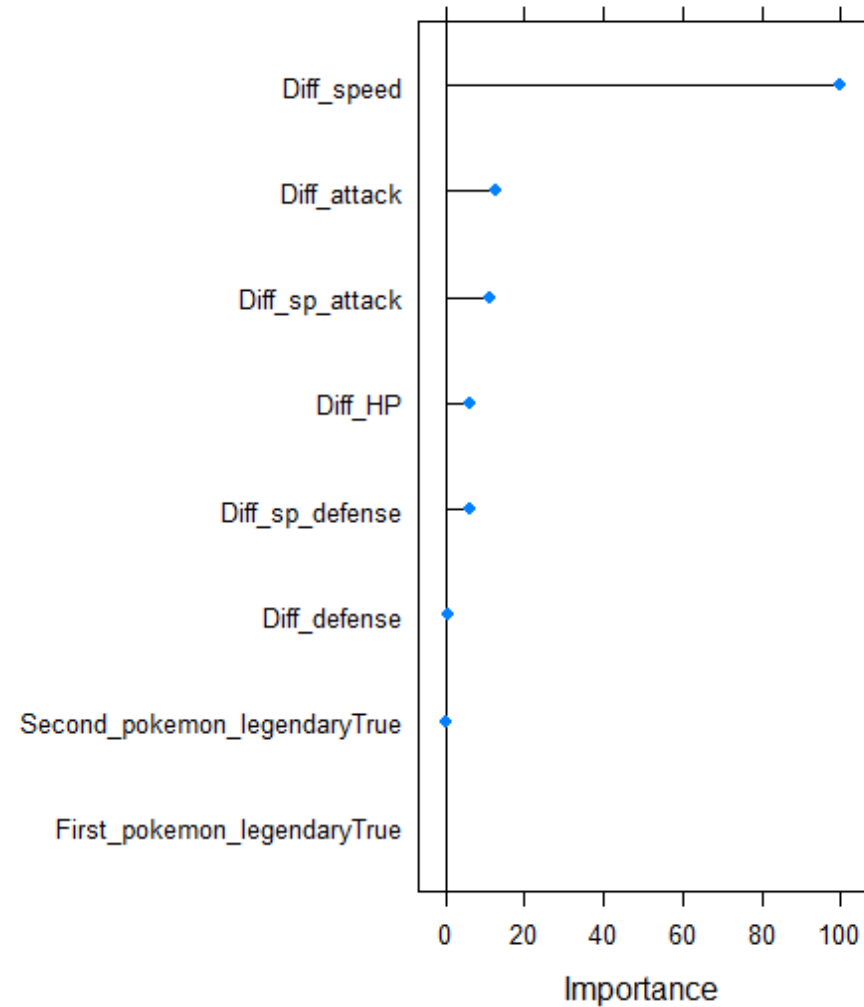


Modello definitivo



Ranking delle feature

- Effettuato il training del modello definitivo, abbiamo provveduto a determinare un ranking sull'importanza delle feature utilizzate ottenendo la seguente rappresentazione
- Risulta evidente come diff_speed sia il parametro più importante su cui basare la classificazione, non a caso si trova come radice del Decision Tree mostrato nella precedente diapositiva



Analisi dei risultati

Dopo aver eseguito il training ed il testing del modello sono state calcolate le seguenti misurazioni:

- **10 fold Cross-validation Confusion Matrix**
- **Prediction Confusion Matrix**
- **ROC Curve**



Risultati

```
Cross-Validated (10 fold) Confusion Matrix
(entries are percentual average cell counts across resamples)

      Reference
Prediction no  yes
no      49.4  2.3
yes     3.4 44.9

Accuracy (average) : 0.9429

> print(precision)
[1] 0.9361616
> print(recall)
[1] 0.9547749
> print(f)
[1] 0.9453767
```

Come è possibile notare, attraverso il modello realizzato è stato possibile ottenere dei valori di accuratezza, precision, recall e f-measure molto elevati, il tutto a testimonianza della bontà del classificatore

```
Confusion Matrix and Statistics

      Reference
Prediction no  yes
no      6599   0
yes      0 5900

      Accuracy : 1
      95% CI : (0.9997, 1)
No Information Rate : 0.528
P-Value [Acc > NIR] : < 2.2e-16

      Kappa : 1
McNemar's Test P-Value : NA

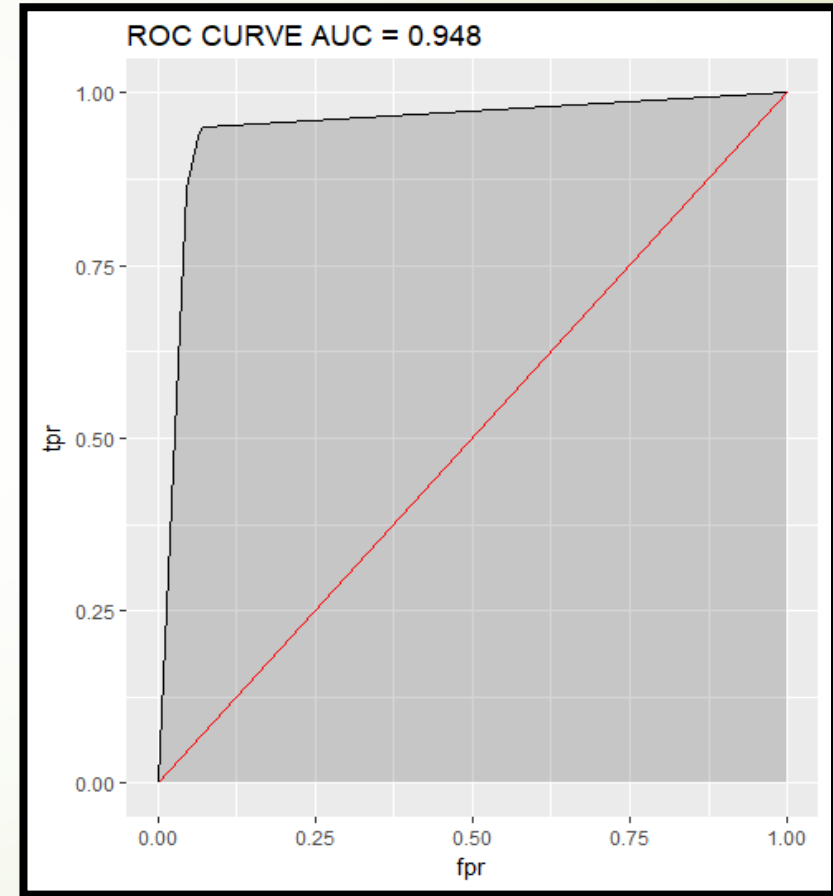
      Sensitivity : 1.000
      Specificity : 1.000
      Pos Pred Value : 1.000
      Neg Pred Value : 1.000
      Prevalence : 0.528
      Detection Rate : 0.528
      Detection Prevalence : 0.528
      Balanced Accuracy : 1.000

      'Positive' Class : no

> print(precision_pred)
[1] 1
> print(recall_pred)
[1] 1
> print(f_pred)
[1] 1
```

Risultati

Anche per quanto riguarda la curva ROC, i valori rilevati dell'area sottostante la curva sono molto elevati e prossimi ad 1, a testimonianza di quanto affermato precedentemente



Predicting Pokémon Fights Outcomes

Marco Belotti (793675)

Francesco Bombarda (794976)

Antonio Vivace (793509)

A.A. 2017-2018

