

# Drug Reviews Dataset

Analisi delle recensioni dei farmaci

Anastasiya Kozemko, Camilla Moretti, Gift Aighobahi,  
Mychael Fokou

# Obiettivo:

- Condurre un'**Analisi Esplorativa dei Dati** (EDA) per comprendere distribuzioni, pattern ricorrenti e relazioni tra le variabili del dataset.
- Applicare tecniche di **Clustering** per identificare gruppi di farmaci con caratteristiche e comportamenti simili.
- Sviluppare un modello di **Machine Learning** capace di prevedere il rating dei pazienti sulla base delle loro esperienze, condizioni trattate e caratteristiche dei farmaci.

# Pulizia del Dataset

# Cella delle conditions

## PROBLEMA

- più malattie in una cella ma separatore non standardizzato



- varianti lessicali associate ad una stessa malattia



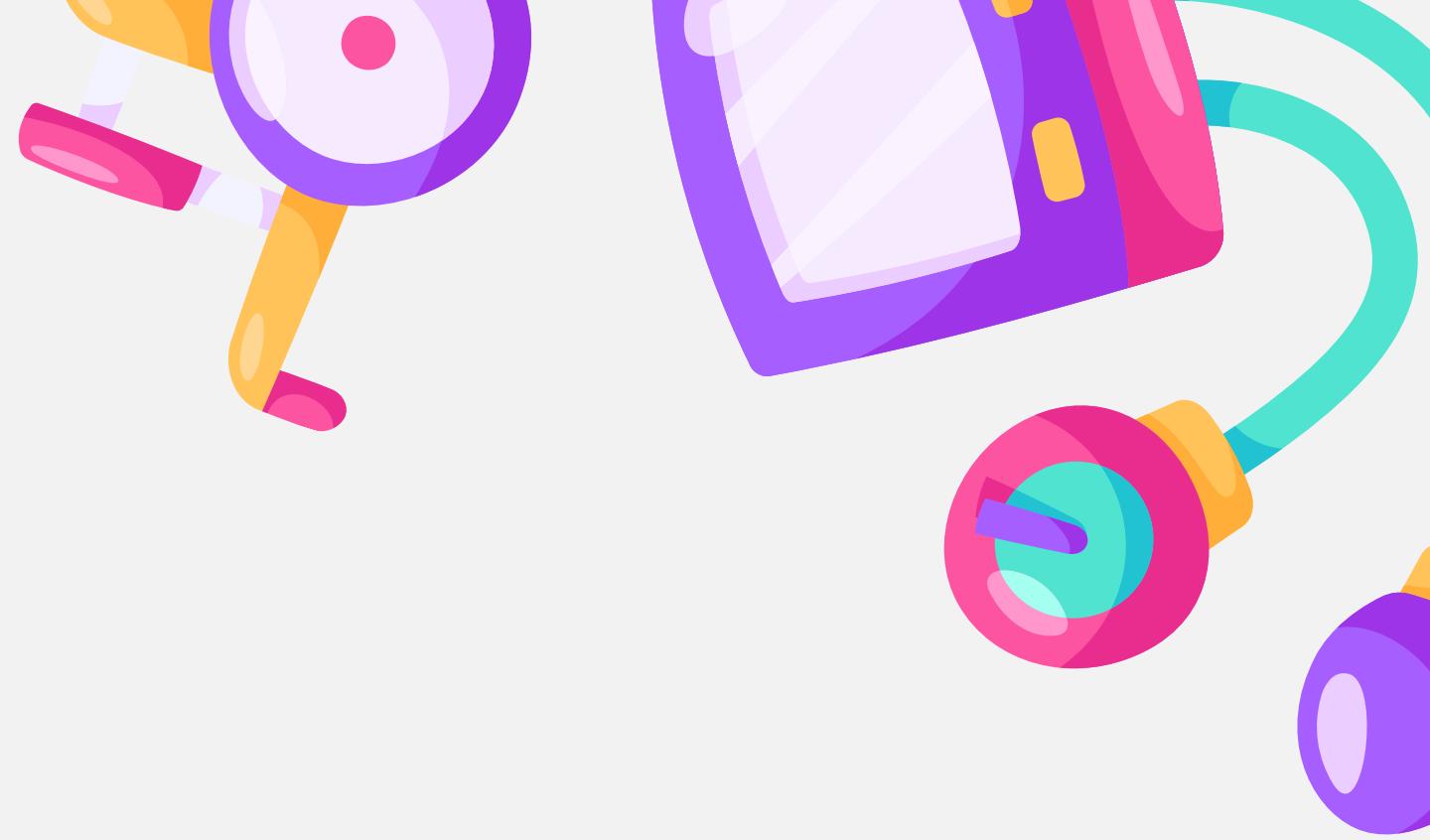
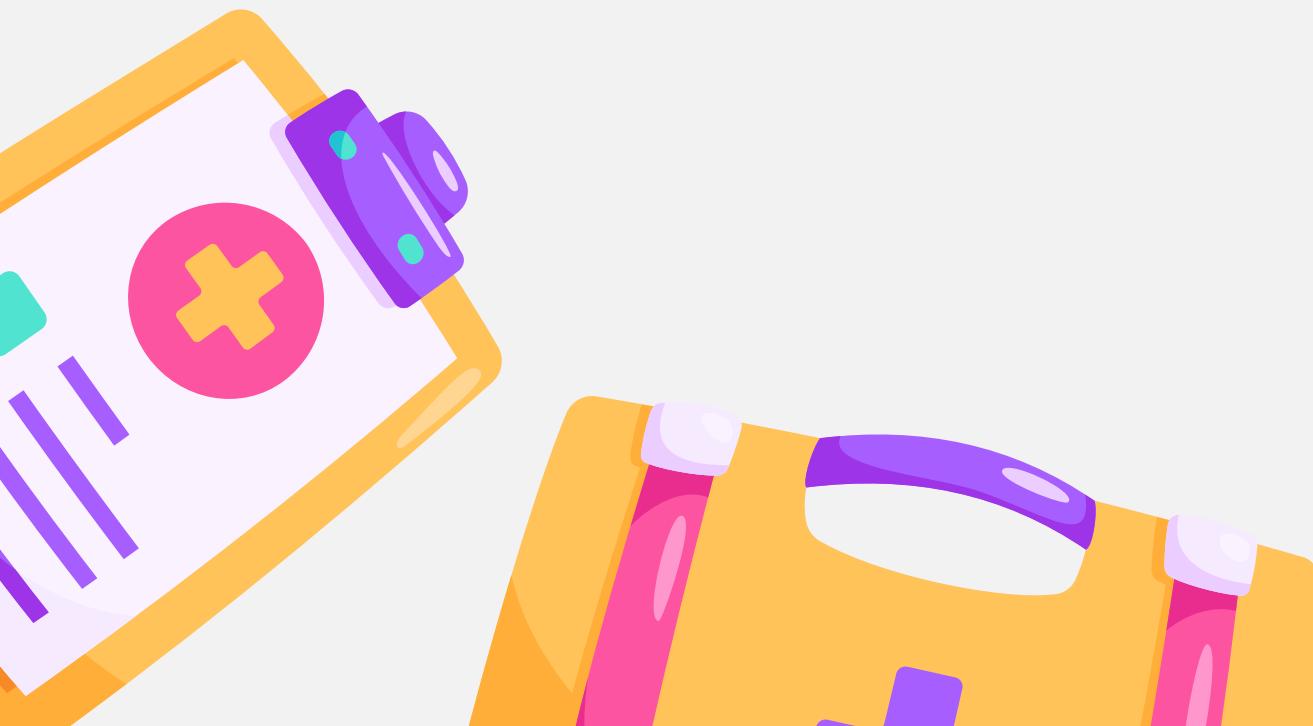
## SOLUZIONE

- creo nuove righe, una per ogni condizione

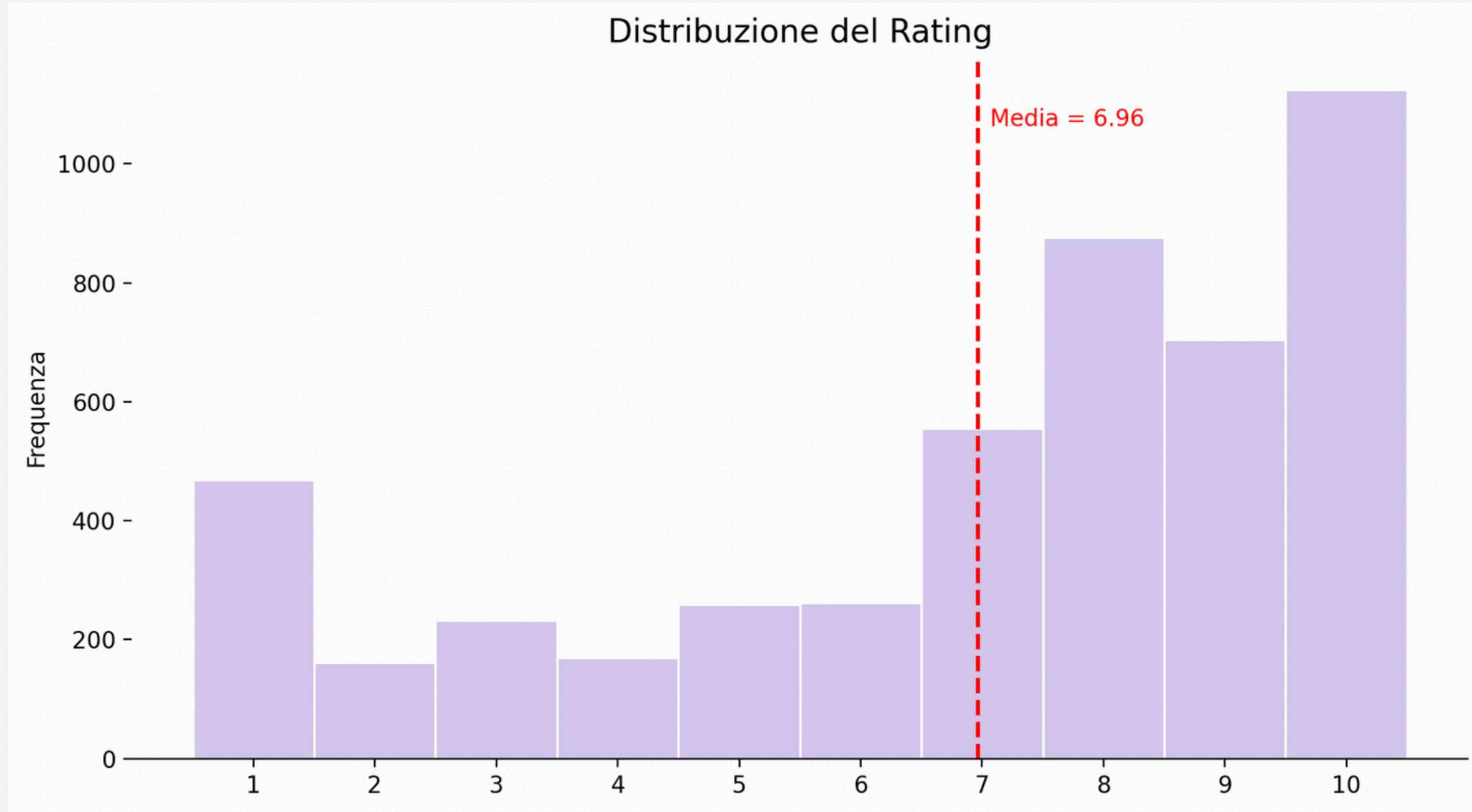
- regex per creare categorie ampie di condizioni simili

- dizionario per identificare eccezioni uniche a bassa frequenza

# Analisi Esplorativa del Dataset



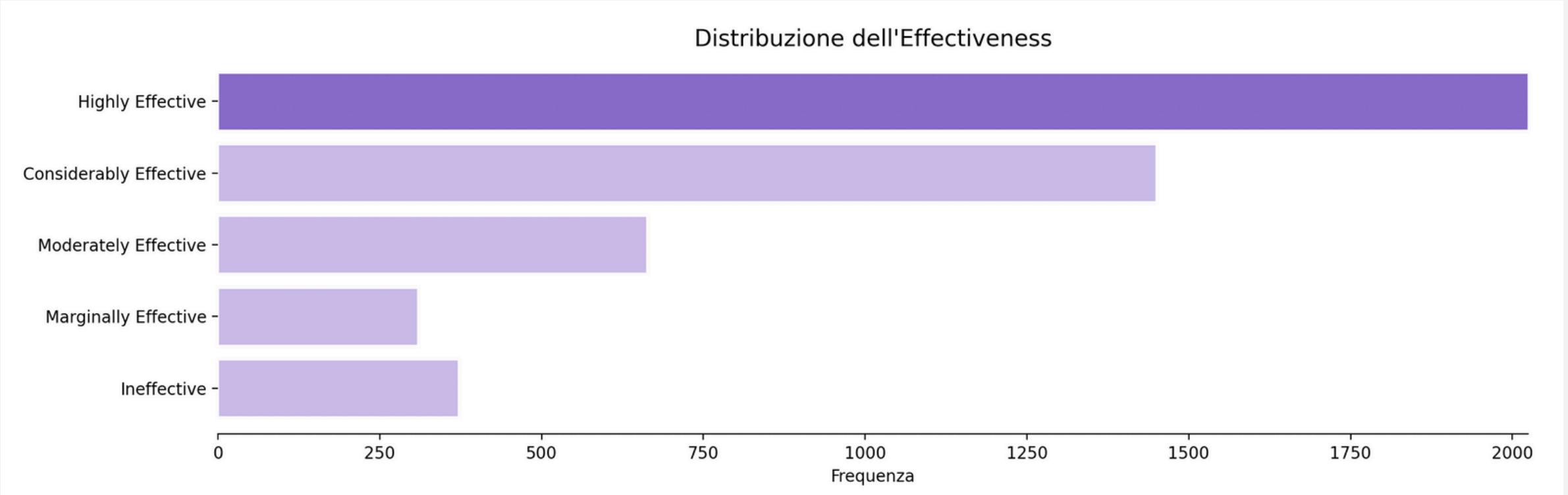
# Distribuzione del Rating del Dataset



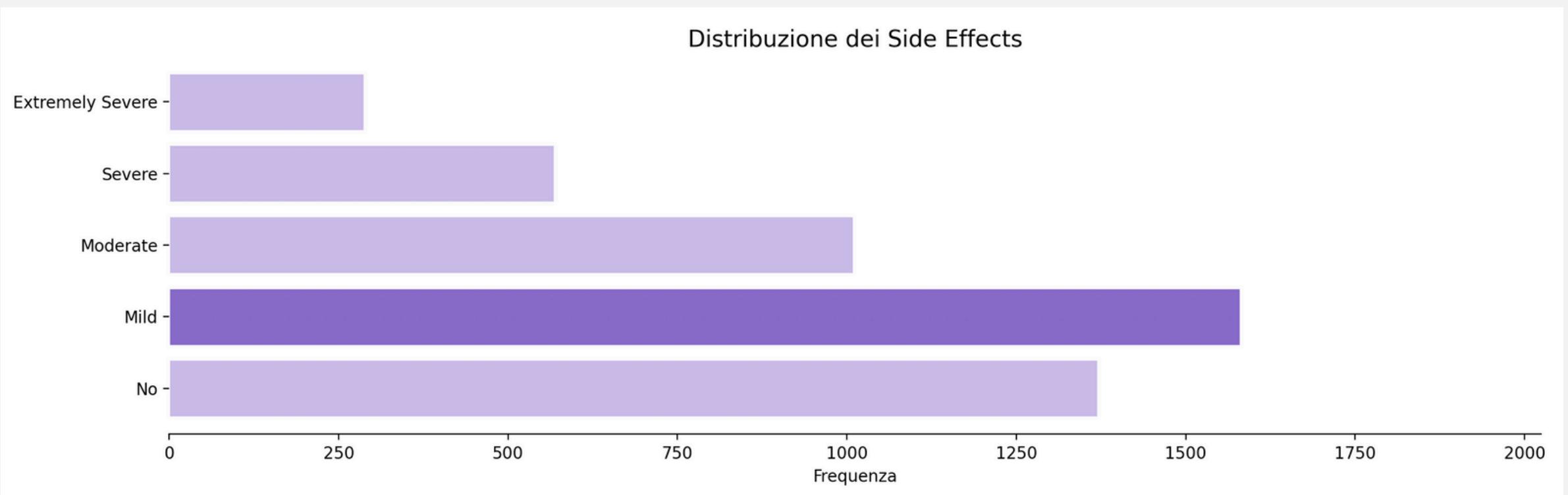
L'istogramma mostra:

- la **distribuzione** dei punteggi assegnati dagli utenti ai farmaci
- la linea rossa tratteggiata indica la **media globale**

# Effectiveness e Side effects

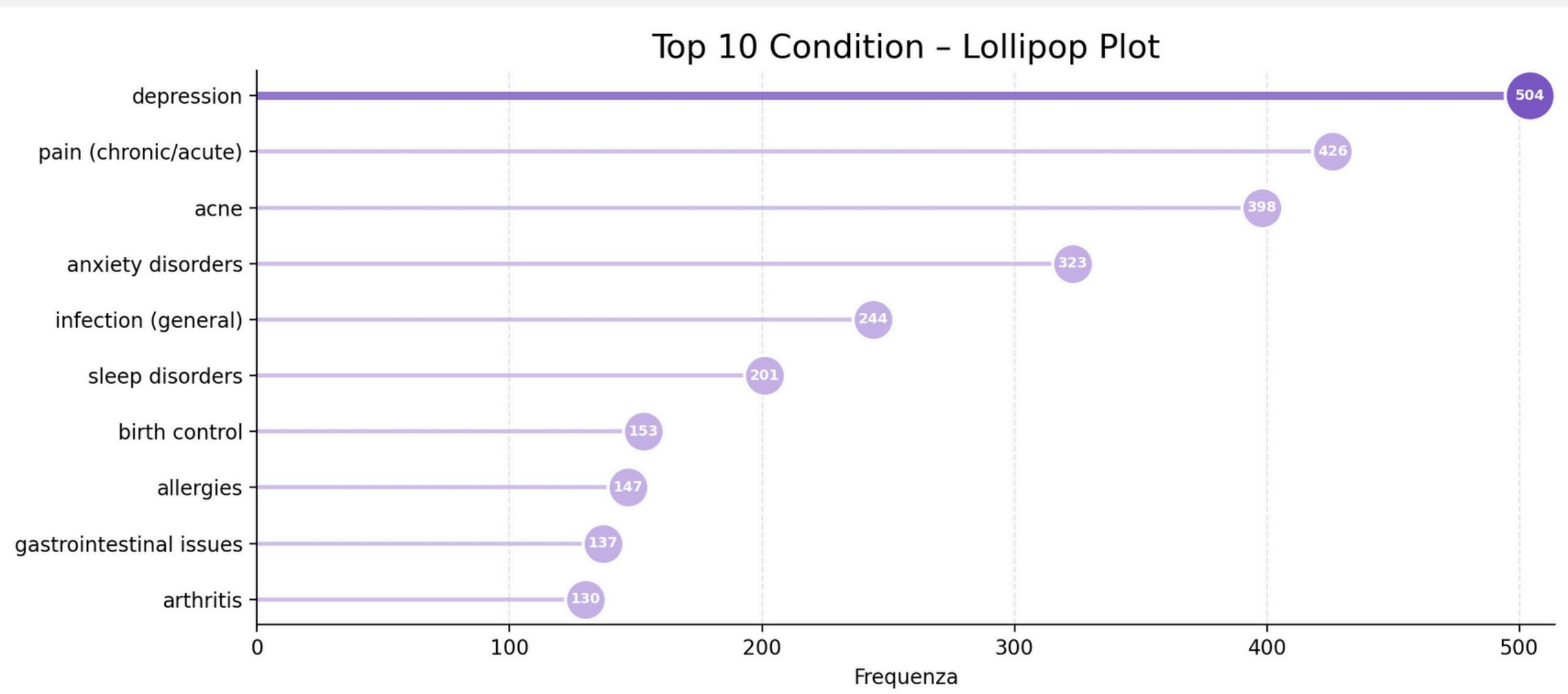


La maggior parte delle osservazioni si concentra su livelli medio-alti di efficacia, indicativo di una percezione generalmente positiva dei farmaci.



Si nota che la categoria “Mild effects” domina, suggerendo che molti trattamenti risultano ben tollerati.

# Le 10 condition più recensite



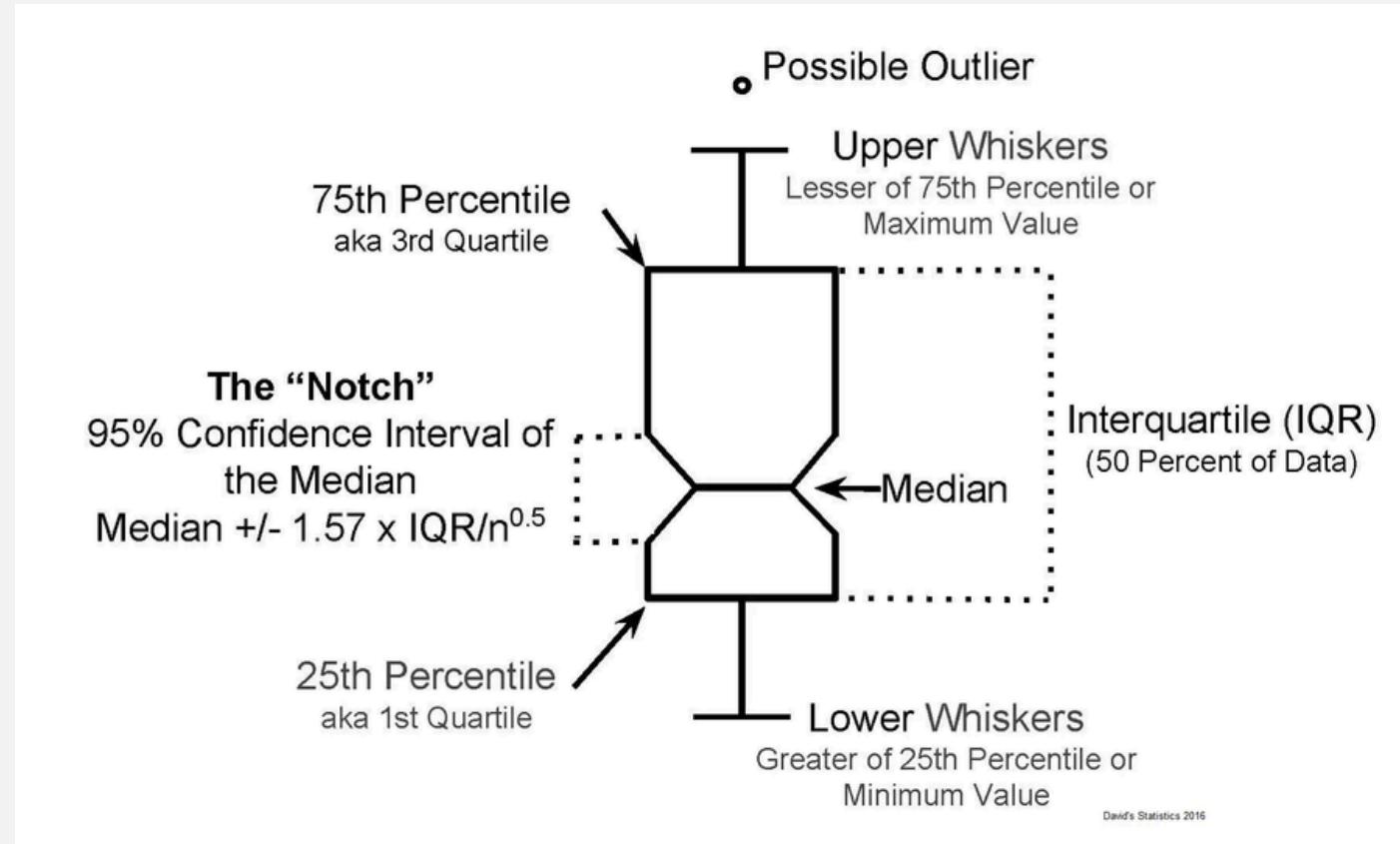
Questo visualizzazione permette di rispondere alle seguenti domande:

- quale condizione è più **frequente**?
- come si **posizionano** le altre rispetto a essa?

Il Lollipop Plot evidenzia in modo elegante la più frequente e permette di confrontare subito le distanze.

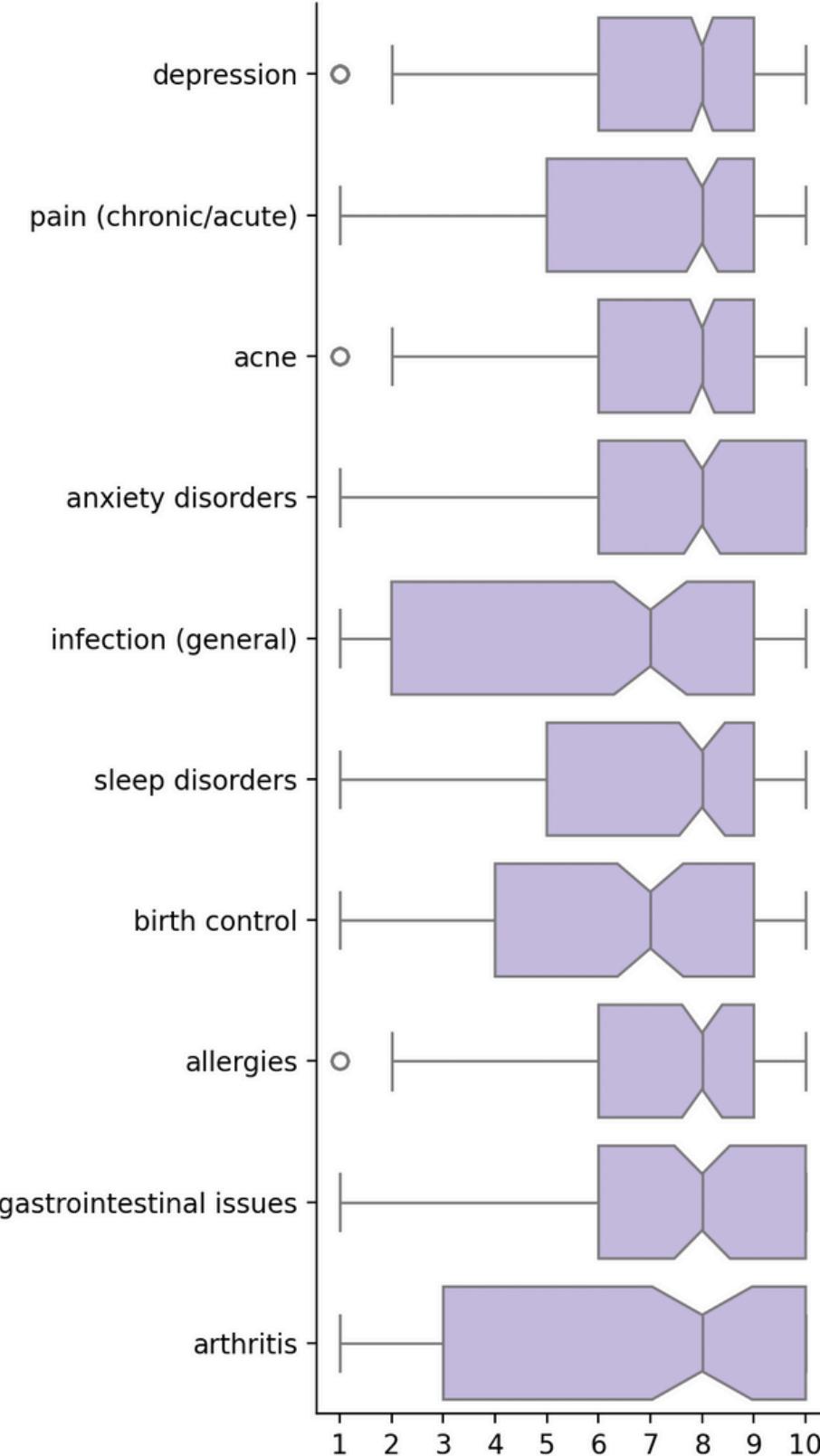


# Rating dei farmaci per le principali condition

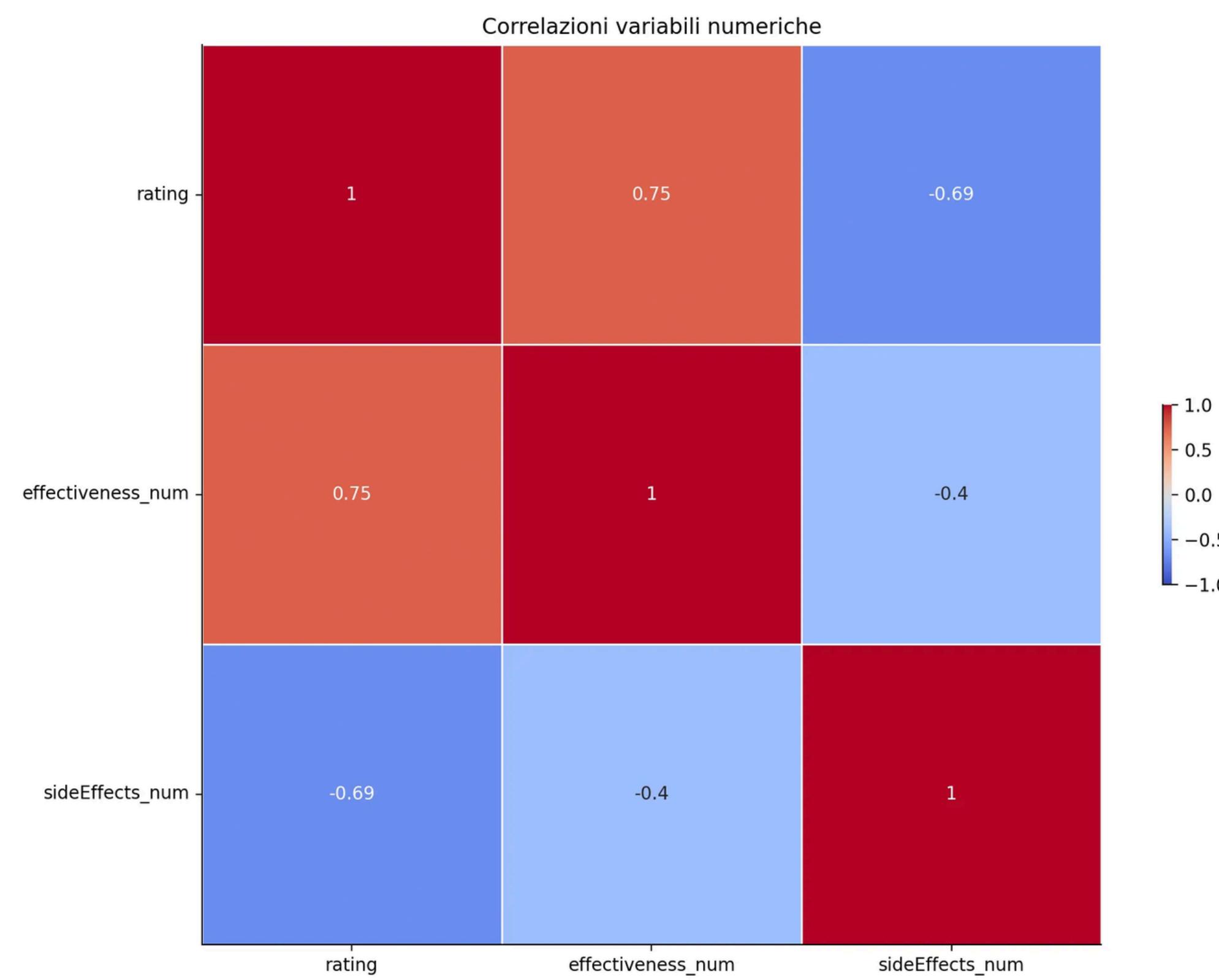


Ci permette di visualizzare come si **distribuisce** il rating assegnato dagli utenti ai farmaci più recensiti per ciascuna condizione, evidenziando mediana, dispersione e possibili outliers.

Distribuzione Rating per le Top 10 Condition Standardized



# Heatmap delle Correlazioni



Viene generata una matrice di correlazione tra:

- `rating` (1–10)
- `effectiveness_num` (1–5)
- `sideEffects_num` (1–5)

Esiste una relazione lineare tra

- Farmaci più efficaci → rating più alto?
- Più effetti collaterali → rating più basso?

📌 + un farmaco è percepito come **efficace** → + alto è il **rating** finale che gli utenti assegnano.

📌 + un farmaco provoca **effetti collaterali** → + il **rating** complessivo scende.

📌 + un farmaco è percepito come **efficace** → - severi sembrano i **side effects**.

# Analisi dei farmaci per una specifica condizione

Poiché dei farmaci hanno **poche recensioni** ma **rating molto alti**, viene usato un **Bayesian mean** per correggere questo effetto (overestimation by low sample size):

**X**

$$[\text{Bayes Mean} = \frac{n}{n+m} \cdot \text{rating medio} + \frac{m}{n+m} \cdot \text{media globale}]$$

- n → nr recensioni del farmaco
- m → parametro di smoothing (10)
- rating medio del farmaco
- media globale della condizione (depression)

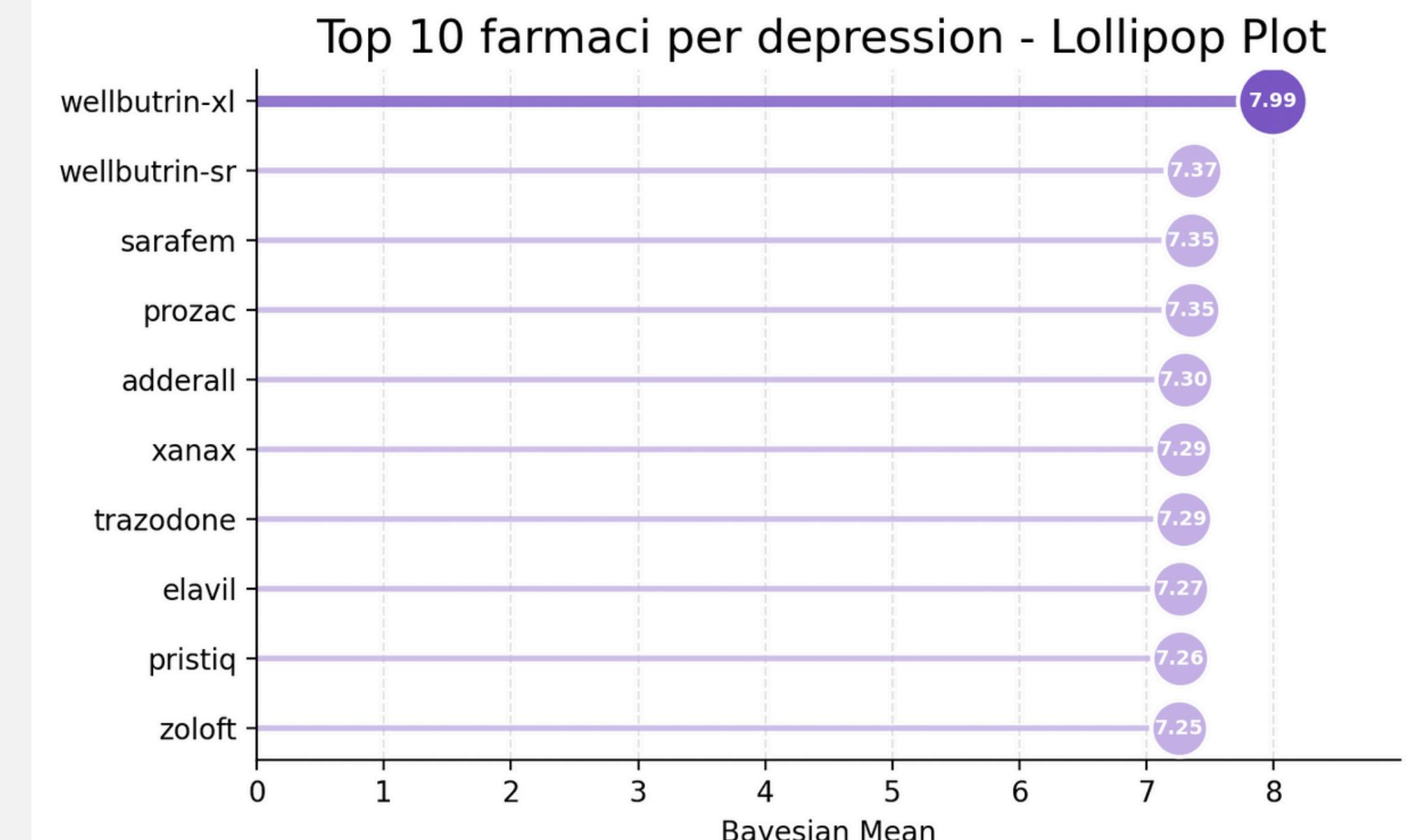
== Calcolo Bayesian Rating ==

Top 10 farmaci (Bayesian Ranking):

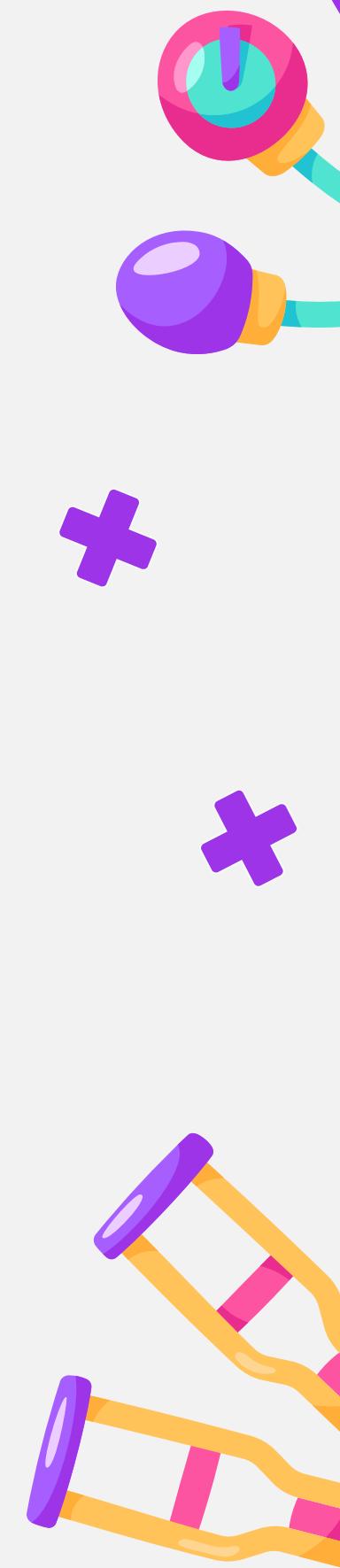
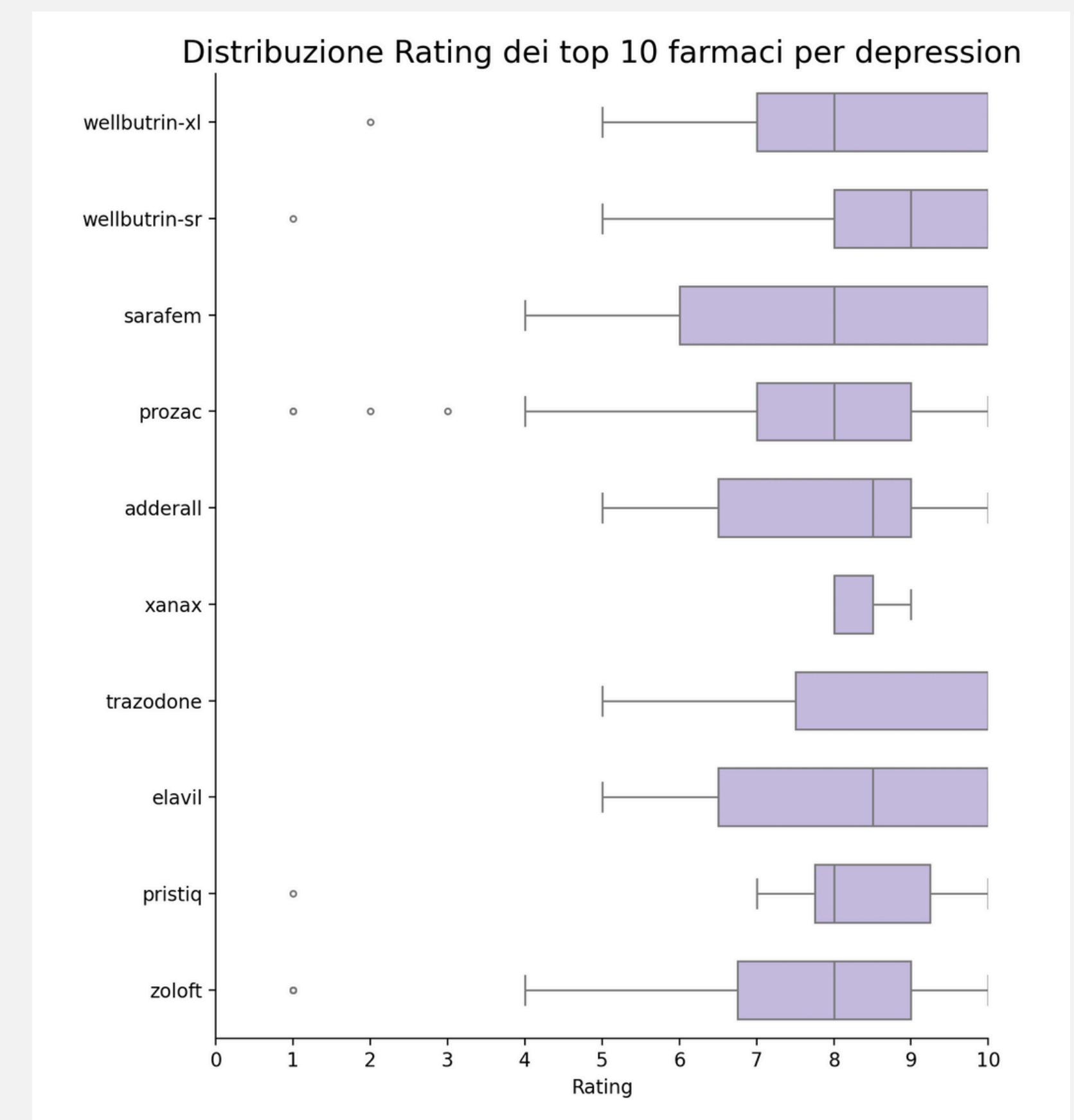
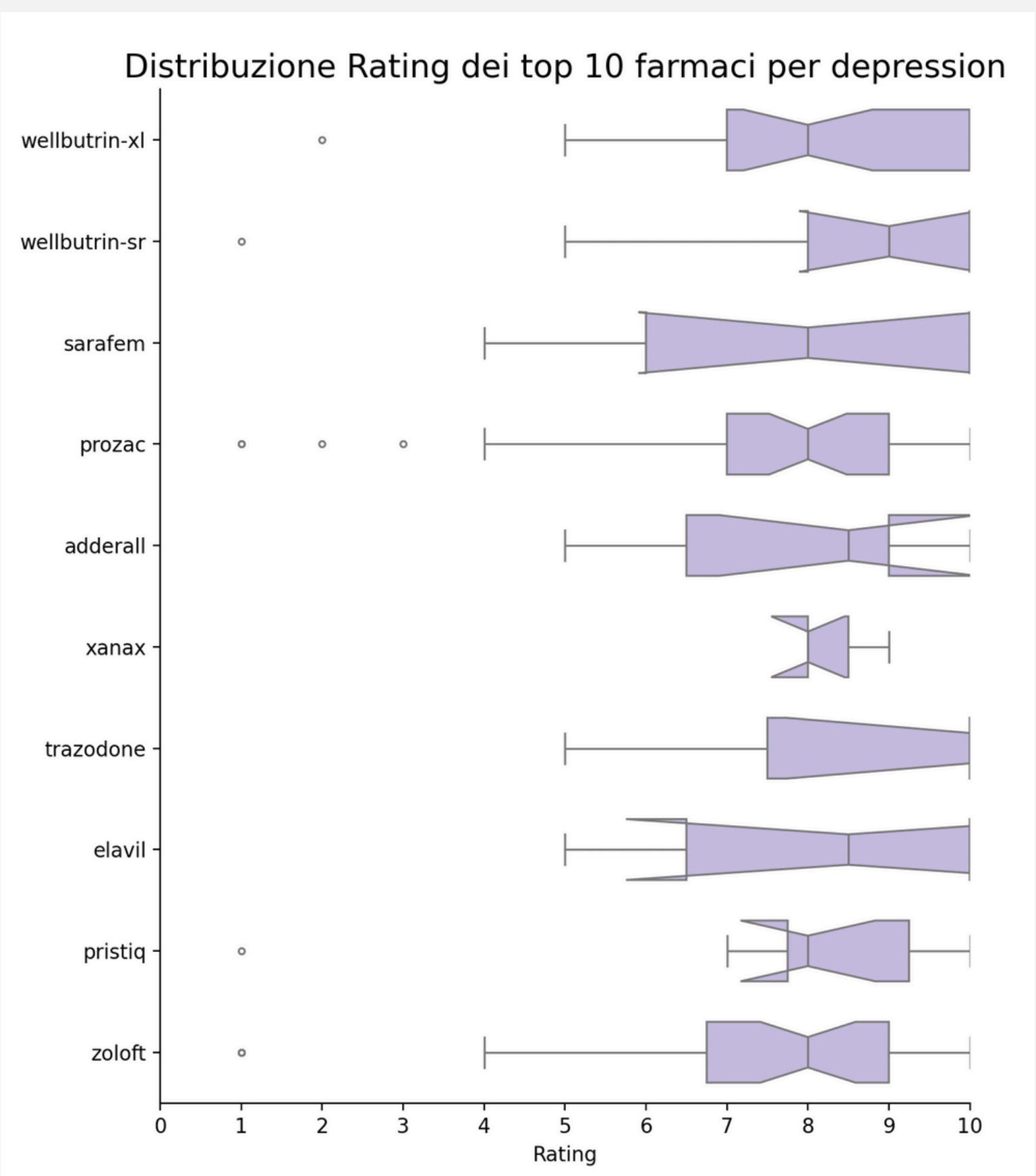
urlDrugName	rating_mean	n_reviews	bayes_mean
wellbutrin-xl	8.285714	35	7.993827
wellbutrin-sr	7.875000	8	7.373457
sarafem	7.777778	9	7.353801
prozac	7.441860	43	7.353249
adderall	7.833333	6	7.295139
xanax	8.333333	3	7.286325
trazodone	8.333333	3	7.286325
elavil	8.000000	4	7.265873
pristiq	7.625000	8	7.262346
zoloft	7.333333	36	7.254831

Farmaci penalizzati (rating alto ma poche recensioni):

urlDrugName	rating_mean	n_reviews	bayes_mean	delta
zyprexa	10.0	1	7.247475	-2.752525
buprenorphine	10.0	1	7.247475	-2.752525
valium	10.0	1	7.247475	-2.752525
prempro	10.0	1	7.247475	-2.752525
tramadol	10.0	1	7.247475	-2.752525

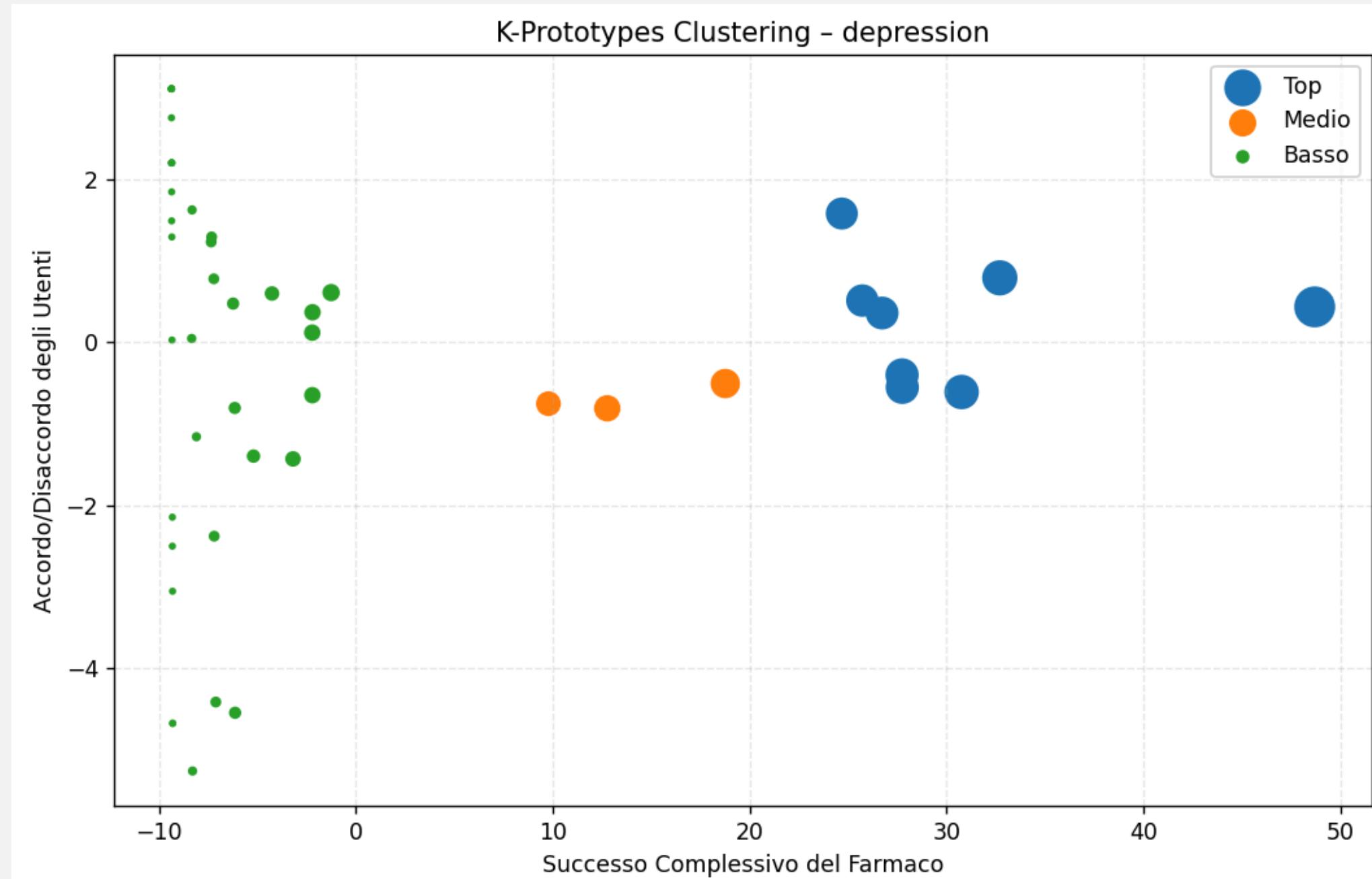


# Analisi dei farmaci per una specifica condizione



# Clustering

# K-Prototypes



Algoritmo creato per la gestione di dataset contenenti feature numeriche e categoriali (dati misti)

Media: per variabili numeriche

Moda: per variabili categoriali

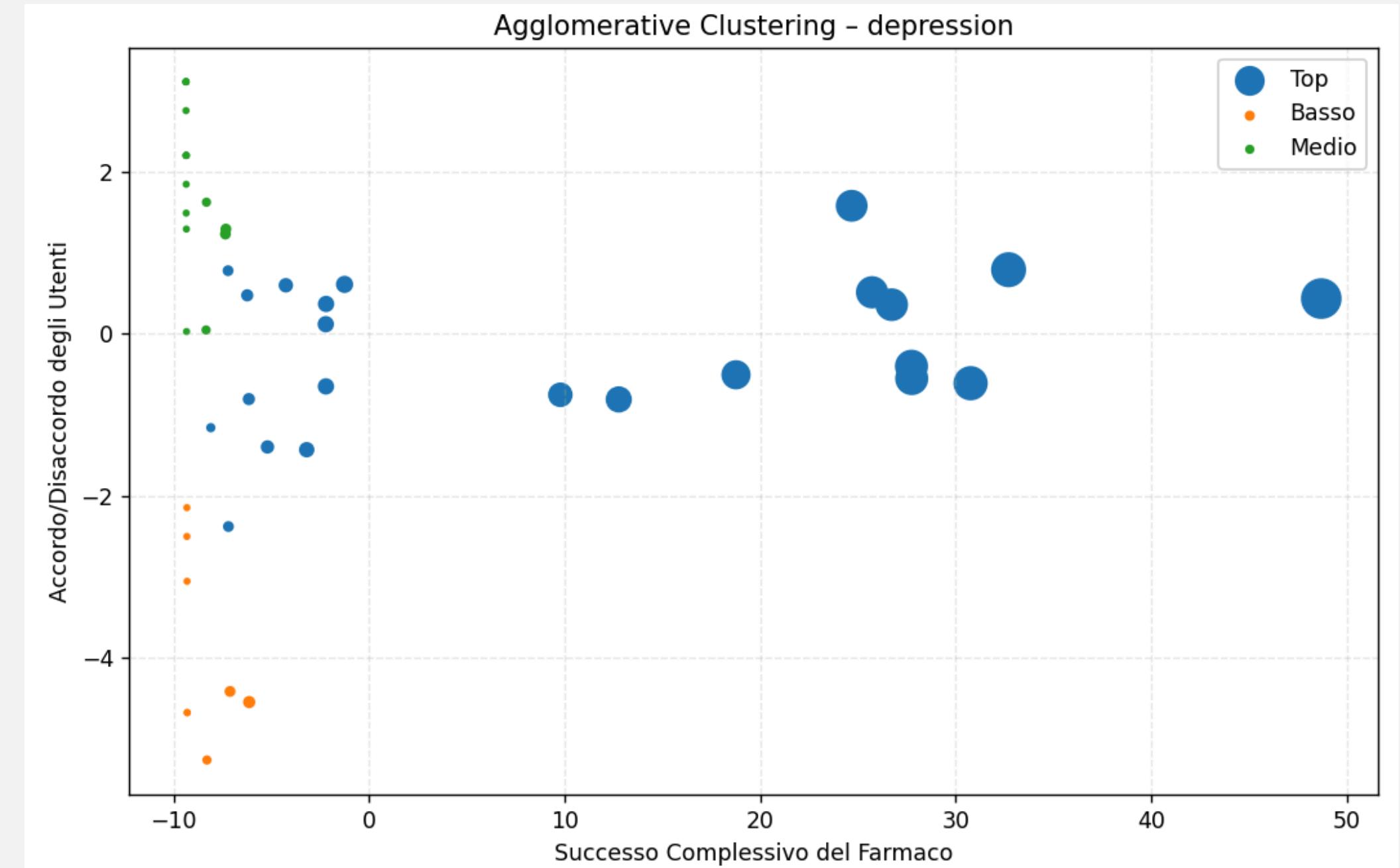
**Criterio di raggruppamento:** mettere insieme dati (farmaci) che hanno vicine le variabili numeriche e simili le variabili categoriche.

# Agglomerative Clustering

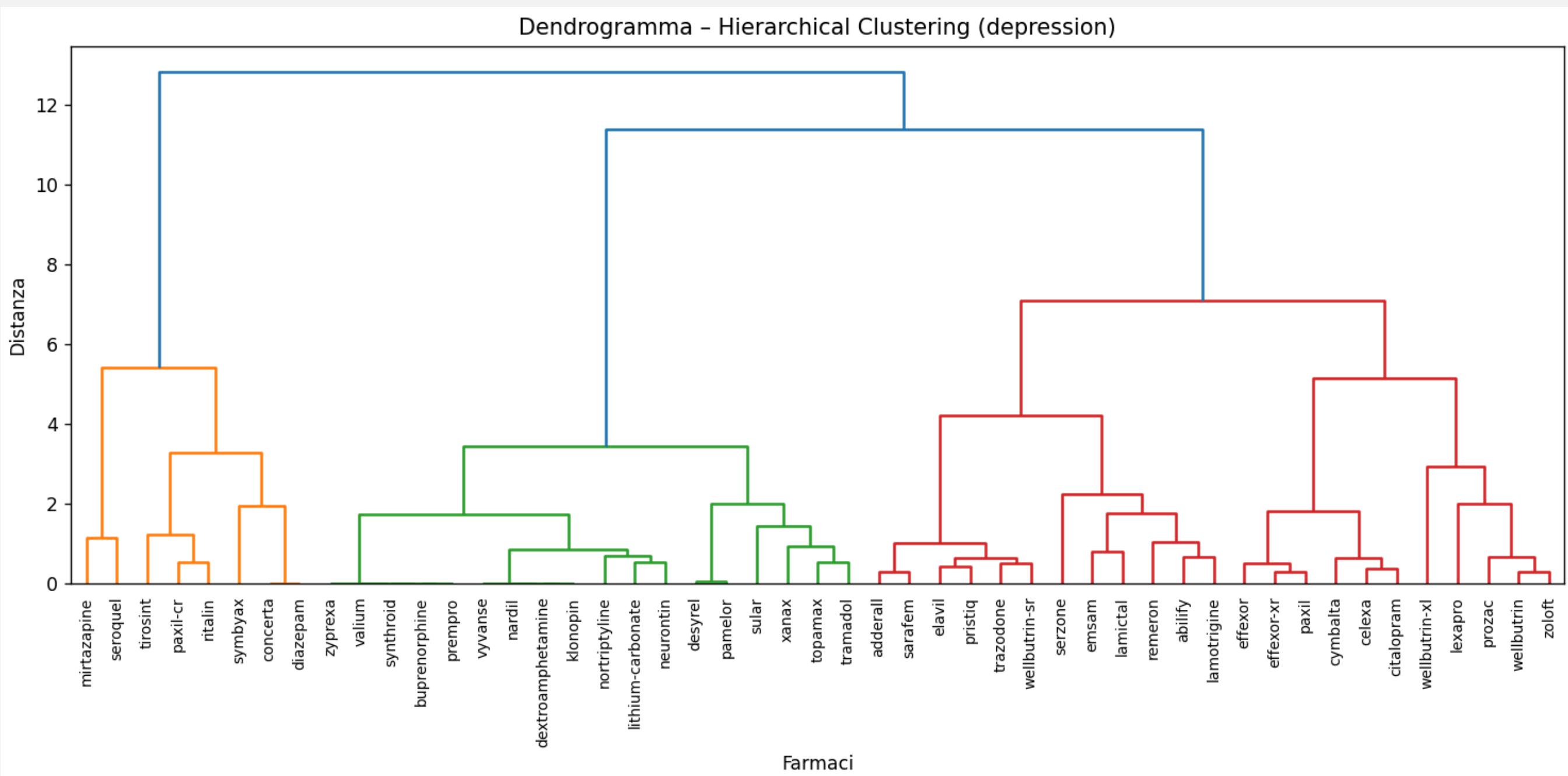
Algoritmo di clustering gerarchico di tipo bottom-up

## Criterio di raggruppamento

1. Calcolo della distanza tra i cluster
2. Fusione progressiva
3. Visualizzazione della gerarchia



# Agglomerative Clustering



Raggruppa i farmaci associati ad una precisa condizione in base alla loro distanza

La **distanza** sull'asse Y indica quanto sono dissimili due farmaci.

# Predizione rating



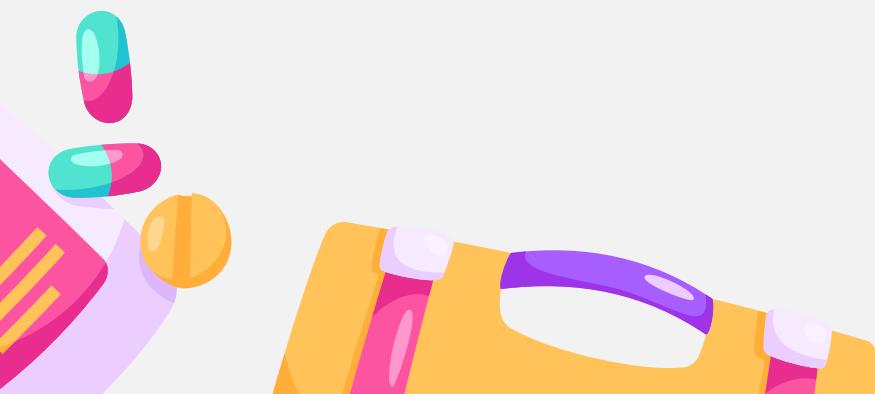
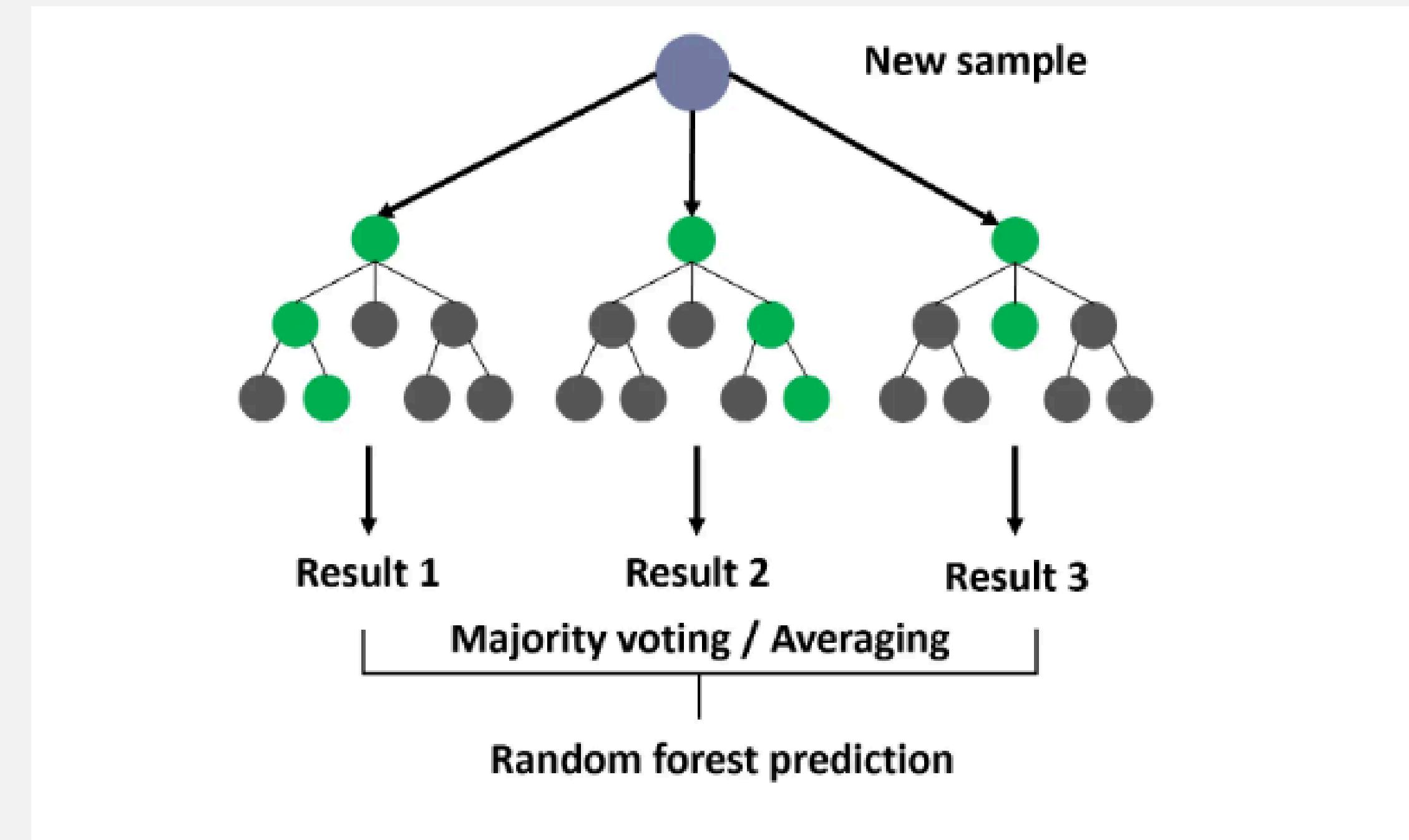
# Previsione rating sui farmaci



Random Forest Regressor

## INPUT

- effectiveness
- side effects
- conditions





# Analisi Comportamentale



CORRELAZIONE DI PEARSON → INDICE DI TOLLERANZA

Si dice **coefficiente di correlazione** delle due variabili  $x$  e  $y$  il numero

$$r = \frac{S_{xy}}{\sqrt{s_x^2 \cdot s_y^2}}$$

dove  $s_x^2$  e  $s_y^2$  sono le **varianze** delle variabili  $x$  e  $y$ .

$S_{xy}$  è la **covarianza** = varianza calcolata mettendo in relazione entrambe le variabili.

Indice di Tolleranza =  
Correlazione(Gravità Effetti Collaterali | Rating di Soddisfazione)





# Analisi Comportamentale

3 PROFILI DI TOLLERANZA → INDICE DI TOLLERANZA

Profilo “Gotta Win”

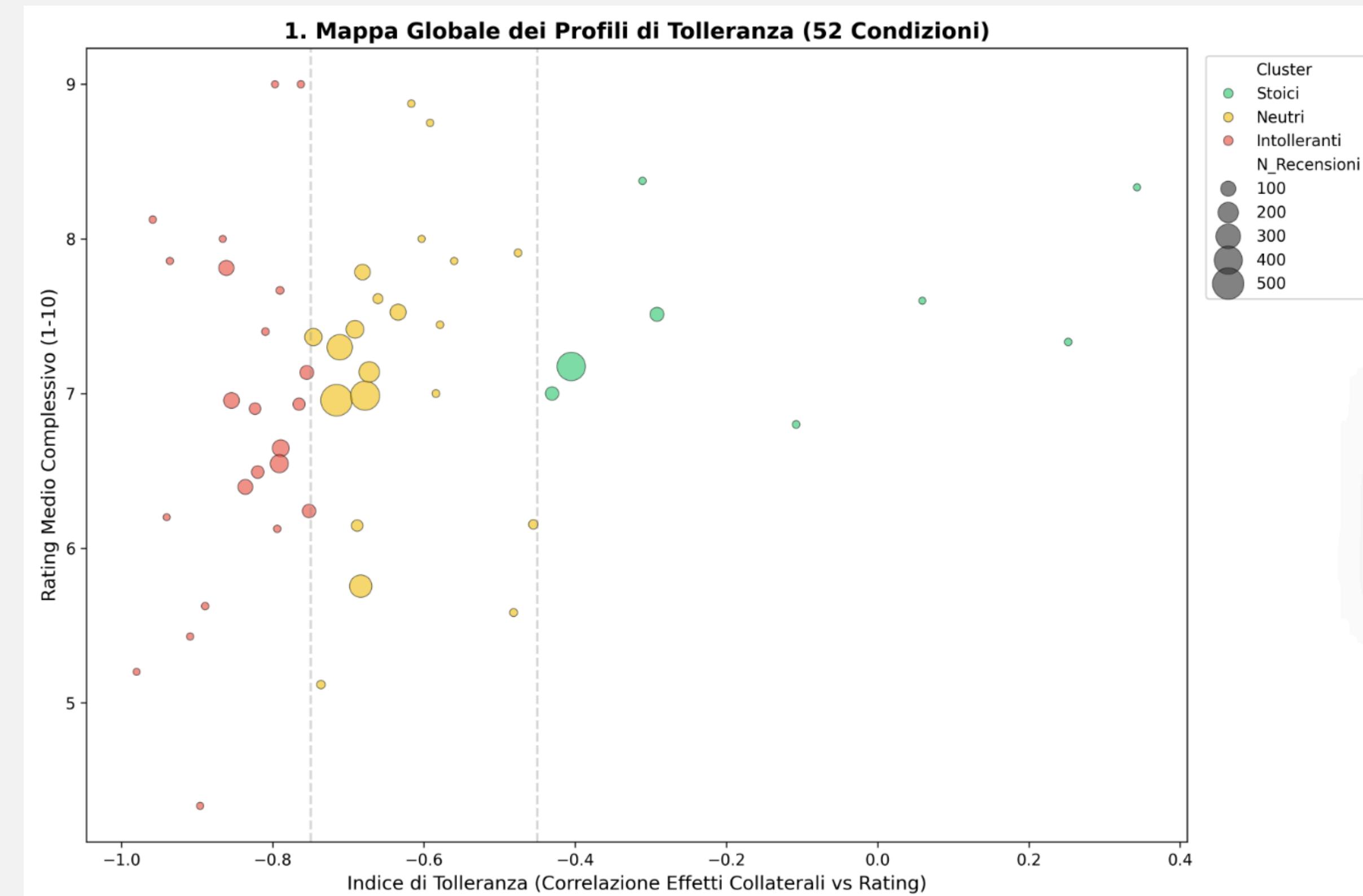
Profilo “No Win”

Zona	Coefficiente	Esempi	Interpretazione
Bassa Tolleranza	[-1, -0.6]	Menopausa, Ipertensione	La malattia crea amgiore intolleranza agli effetti collaterali
Media Tolleranza	(Tra -0.6, -0.4)	ADHD, Dolore Cronico	il voto dei pazienti è influenzato da entrambi i fattori in modo bilanciato
Alta Tolleranza	(-0.4, 0)	Acne, Rosacea	L'effetto collaterale è un fastidio minore, il voto si basa sull'efficacia.



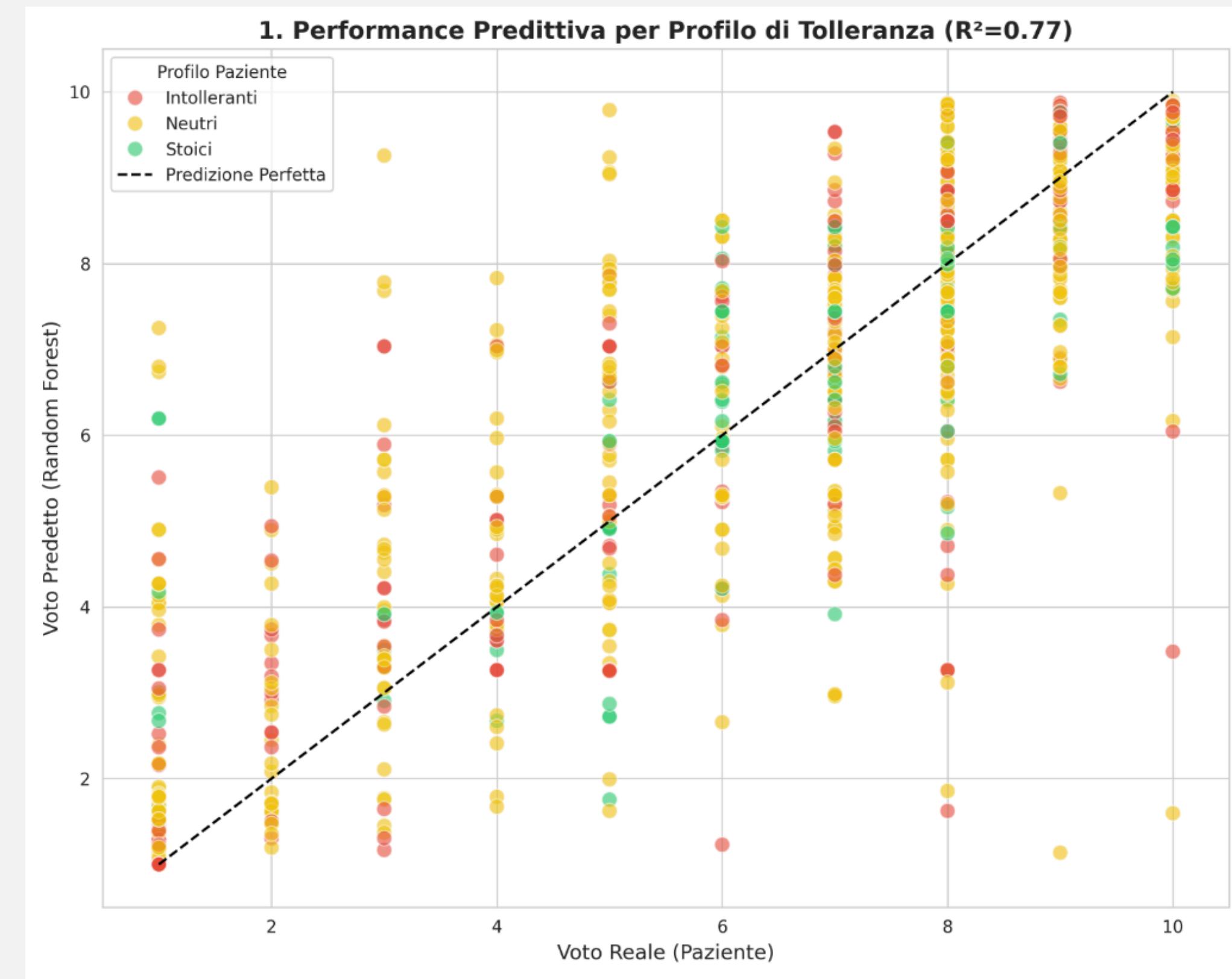
# Analisi Comportamentale

## SCATTERPLOT DELL'INDICE DI TOLLERANZA x RATING



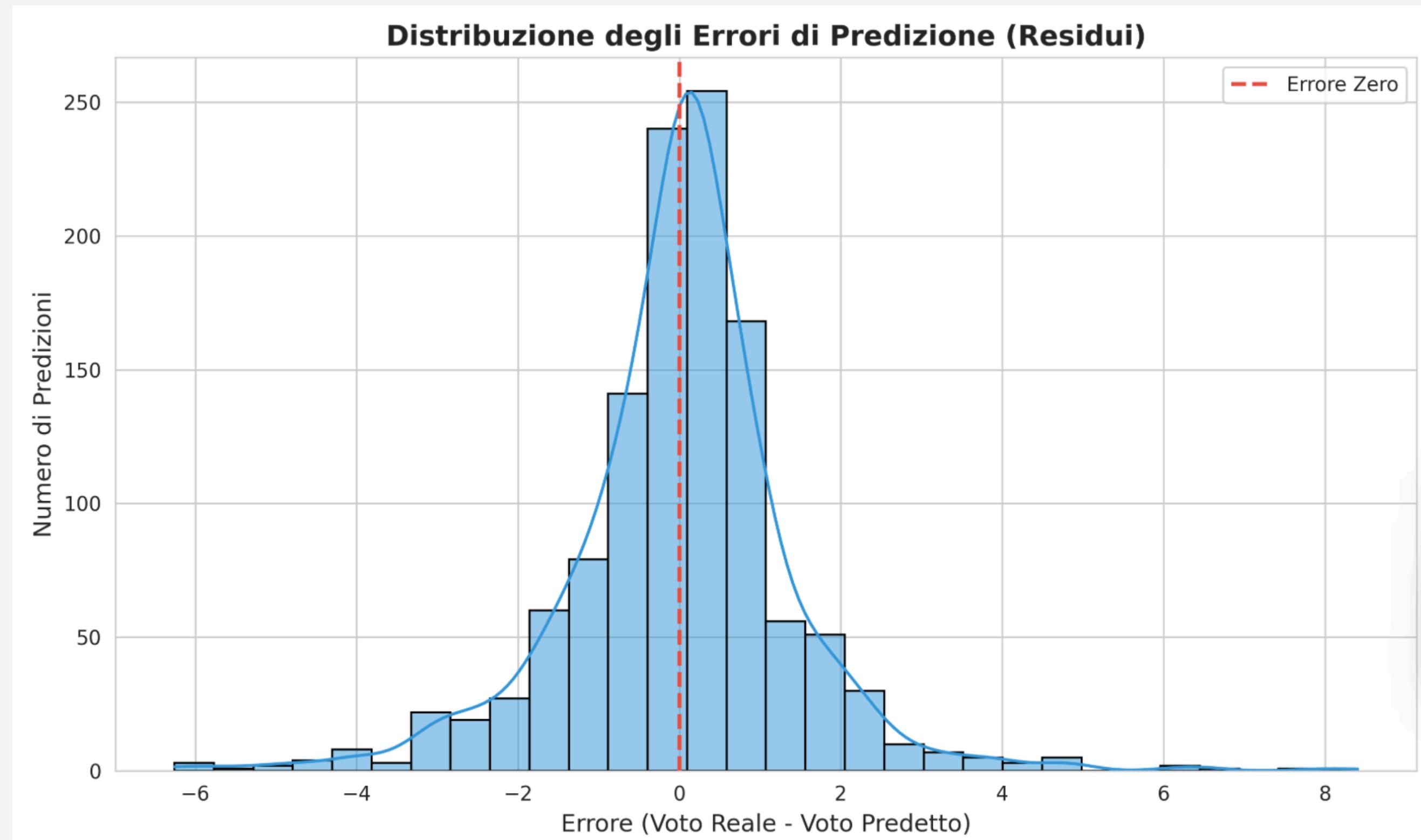
# Output: analisi performance

Tabella delle predizioni → Scatter plot



# Output: analisi performance

## ISTOGRAMMA





# Output: analisi performance

Affidabilità:  $R^2$  (R-Quadro) = 77%

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} = 1 - \frac{\text{Varianza Non Spiegata}}{\text{Varianza Totale}}$$

- $\sum(y_i - \hat{y}_i)^2$ : Errore del tuo modello (la somma degli errori quadratici).
- $\sum(y_i - \bar{y})^2$ : Errore del modello *naive* (la varianza totale dei dati, dove  $\bar{y}$  è il Rating Medio).

$$R^2 = 1 - \frac{\sum(\text{Rating Reale} - \text{Rating Predetto})^2}{\sum(\text{Rating Reale} - \text{Rating Medio})^2}$$

$y_i$ : Il **Rating Reale** (il voto vero del paziente) per l'osservazione  $i$ .

$\hat{y}_i$ : Il **Rating Predetto** dal tuo modello Random Forest per l'osservazione  $i$ .

$\bar{y}$ : Il **Rating Medio** (la media aritmetica) di tutti i Rating Reali  $y_i$ .

# Output: analisi performance

RMSE (Root Mean Squared Error) = 1.41  
in una scala da 0 a 10

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$$

- $N$ : Numero totale di recensioni nel set di test.
- $y_i$ : Il **Rating Reale** (il voto vero del paziente).
- $\hat{y}_i$ : Il **Rating Predetto** dal tuo modello Random Forest.

# Output: analisi performance

## FEATURE OF IMPORTANCE

- Effectiveness: 61%
- Side effects: 30%
- Condition: 9%

### REGOLA DI SPLIT

criterio che l'albero decisionale utilizza ad ogni nodo per giungere al rating predittivo  
(complessivo di tutte le decisioni)

### CRITERIO DI SPLIT

Random Forest premia la **feature** (parametro in input) che "ordina" meglio i dati, ovvero che restituisce (nel nostro caso) 100 predizioni dai valori più vicini possibile

### DECREMENTO TOTALE

Ogni regola di split influenza la varianza sulle predizioni tra gli alberi, questa influenza viene chiamata "**guadagno**" ed esprime quanto si avvicina la random forest ad un valore "**puro**", ovvero con i rating ordinati (tutti vicini) → **questo guadagno è misurato in una scala percentuale per ogni feature**



Thank You

