



# The Joe Rogan Experience





# Per chi non lo conoscesse:

Il Joe Rogan Experience è un **podcast** condotto da **Joe Rogan**, noto comico e presentatore.

Lanciato nel **2009**, è diventato **uno dei podcast più seguiti al mondo**. Rogan conversa con **ospiti di vari settori**, tra cui scienza, politica, sport, cultura pop e filosofia.

Gli episodi sono spesso lunghi e trattano argomenti in modo profondo, **senza censura**, offrendo uno spazio per **discussioni autentiche e senza freni**.



# Qualche numero:

Il motivo principale per cui ho scelto joe rogan è la sua **popolarità**, il canale conta quasi **20 mln** di iscritti e oltre **6 mld** di visualizzazioni.

Gli episodi del podcast sono più di **2000** mentre gli **ospiti** differenti sono circa **1000**.

L'idea è quella di sfruttare la grande quantità di episodi e persone diverse per trovare delle **correlazioni tra la performance dei video e le caratteristiche degli ospiti**.



# Obiettivo:

L'obiettivo del progetto è sviluppare uno **strumento in grado di prevedere la performance di un futuro episodio** del podcast, sulla base delle caratteristiche dell'ospite.

Il progetto si articola in **due fasi**:

1. Modello **base**: regressione utilizzando esclusivamente i dati relativi all'ospite.
2. Modello **avanzato**: regressione arricchita con informazioni derivanti da un grafo costruito sulle relazioni tra gli ospiti.



# Raccolta dati:

- **YouTube API**: utilizzata per estrarre informazioni sui video del podcast, tra cui visualizzazioni, data di pubblicazione, titolo, descrizione, ecc...
- **Gemini API**: impiegata per ottenere le caratteristiche degli ospiti, nel particolare: ‘Nazionalità’, ‘Età’, ‘Sesso’, ‘Professione’, ‘Notorietà’ e ‘Numero di follower’.



# Gemini data - 1

Dalla **API di youtube** ho ottenuto il titolo di ogni video poi con uno script ho isolato tutti i nomi degli ospiti:

```
video_id,title,guest,published_at,views,likes,comments,description
HUBAdmcJeNw,Joe Rogan Experience #1794 - Monty Franklin,Monty Franklin,2024-06-27T17:48:42Z,30000,wu6USAUpGDg,Joe Rogan Experience #1793 - Mike Baker,Mike Baker,2024-06-27T17:48:36Z,8618,129,rMNW-orIpU8,Joe Rogan Experience #1792 - Daryl Davis & Bill Ottman,Daryl Davis & Bill Ottman,_cPD7l1LEDA,Joe Rogan Experience #1791 - Sadhguru,Sadhguru,2024-06-27T17:48:26Z,2067001,42409,II41I5KzJAk,Joe Rogan Experience #1790 - Nims Purja,Nims Purja,2024-06-27T17:48:20Z,53188,1477y0P1BGH7cq,Joe Rogan Experience #1789 - Tom Papa,Tom Papa,2024-06-27T17:48:15Z,8978,115,30,"Comedia"
```

HUBAdmcJeNw, Monty Franklin, "Monty Franklin"  
wu6USAUpGDg, Mike Baker, "Mike Baker"  
rMNW-orIpU8, Daryl Davis, "Daryl Davis"  
rMNW-orIpU8, Bill Ottman, "Bill Ottman"  
Create Jira Issue  
\_cPD7l1LEDA, Sadhguru, "Sadhguru"  
II41I5KzJAk, Nims Purja, "Nims Purja"  
7y0P1BGH7cq, Tom Papa, "Comedia"



# Gemini data - 2

Con la lista dei nomi ho sfruttato l'**API di gemini** (gemini-2.5-preview-06-05) per trovare le caratteristiche di ogni ospite:

```
prompt= f"""\n    Fornisci informazioni di {nome} (ospite del podcast di Joe Rogan) in formato JSON.\n\n    Informazioni aggiuntive dall'episodio (potrebbero non essere disponibili):\n    Link del video: https://www.youtube.com/watch?v={video\_id}\n    Descrizione del video: {description}\n\n    Il JSON deve avere le seguenti 'chiavi': (possibili valori) =\n    'Nome': (nome fornito),\n    'Nazionalità': (es. Europa, America, Asia.),\n    'Sesso': (M, F),\n    'Età': (#),\n    'Professione': La più importante. È fondamentale standardizzarla. Ad esempio, invece di avere "Sviluppatore",\n    'Notorietà': Creare una scala consistente. Esempio di scala: Esperto di Nicchia: Noto solo nel suo campo,\n    'Follower social': Il numero di follower sulla sua piattaforma principale.\n    'Topic video': Argomento Principale del Video (Categorica). Di cosa si è parlato? (es. "AI Generativa",\n\n    - Usa il valore JSON `null` per le informazioni che non riesci a trovare dopo una ricerca approfondita.\n    - Considera le informazioni dalla descrizione dell'episodio per contestualizzare meglio la persona.\n    - Rispondi SOLO con il JSON valido, senza testo introduttivo o conclusivo, nemmeno ```json ```. """
```

video\_id, Nome, Nazionalità, Sesso, Età, Professione, Notorietà, Numero di follower social, Topic video  
nNxzn-vRodY, Ron White, America, M, 67, Commedia/Intrattenimento, Personaggio Pubblico, 4100000, Commedia/Storie di vita  
NRVEkc9lxH0, Ben Lamm, America, M, 40, Imprenditoria / Business, Influencer di Settore, 35700, Biotecnologia / De-estinzione  
-9shy1j8wjM, Kyle Dunnigan, America, M, 53, Intrattenimento/Media, Personaggio Pubblico, 952000, "Commedia, attualità e imitazioni"  
EmNE6yNxruc, Dave Smith, America, M, 42, Commedia/Opinionista Politico, Influencer di Settore, 480000, Politica/Attualità  
bJbKgC\_bMpI, Kurt Metzger, America, M, 47, Intrattenimento/Commedia, Influencer di Settore, 198000, "Commedia, Attualità, Cultura"  
bMsJHjt55Do, Francis Foster, Europa, M, 43, Commedia/Intrattenimento, Influencer di Settore, 1100000, Politica/Cultura  
bMsJHjt55Do, Konstantin Kisin, Europa, M, 41, Media/Commentatore Politico, Influencer di Settore, 1100000, Critica Sociale e Politica  
TFLLmFpomLM, Big Jay Oakerson, America, M, 46, Intrattenimento/Commedia, Influencer di Settore, 520000, Commedia e Carriera  
ZQHIDX1lC3k, Scott Payne, America, M, 45, Forze dell'Ordine/Sicurezza, Esperto di Nicchia, 9000, Terrorismo Domestico/Infiltrazione FBI  
207WIA\_BjqI, Dr. Suzanne Humphries, America, F, 60, Medicina alternativa/Ricerca, Influencer di Settore, 215000, Critica ai Vaccini e Immunità  
MjhXtJB\_ZbU, Chris Williamson, Europa, M, 36, Media/Podcasting, Influencer di Settore, 2400000, Crescita Personale/Dinamiche Sociali  
xsqbdBidWE, Josh Waitzkin, America, M, 47, Crescita Personale / Atleta, Personaggio Pubblico, 2000, Apprendimento e Maestria  
FjaGb87pjMc, Bert Kreischer, America, M, 51, Commedia/Intrattenimento, Personaggio Pubblico, 3100000,  
d4m1M0SI5n8, Michael Kosta, America, M, 44, Intrattenimento/Media, Personaggio Pubblico, 104000, Commedia e Sport  
DE9oFxGoMvE, Darryl Cooper, America, M, 35, Podcaster/Storico, Influencer di Settore, 368000, Storia - American History  
vZEcVY2iZsk, Jacques Vallée, Europa, M, 84, Ricerca/Scienza, Personaggio Pubblico, 100, Ufologia/Fenomeni Misteriosi  
Mx8813RLUtc, Josh Dubin, America, M, Legale/Attivismo, Influencer di Settore, 23900, Riforma della Giustizia Penale  
Mx8813RLUtc.J.D. Tomlinson, America, M, 41, Legale/Giustizia, Esperto di Nicchia.. Giustizia e Riforma Legale



# Gemini data - 3

- Gemini è riuscito a fornire quasi tutti i valori con buona precisione. Interventi manuali minimi (età e follower, soprattutto per persone non famose).
- **Categorizzazione** professioni e topic:
  - Output iniziale:
    - ~300 professioni distinte
    - ~1000 topic diversi
  - Ristrutturazione:
    - 9 macro-categorie per le professioni
    - 20 categorie tematiche per i topic



# Gemini data - 4

Per ogni episodio ho associato:

→ Visualizzazioni, like, commenti recuperati in precedenza.

**Normalizzazione** dei dati:

- Per garantire coerenza e confrontabilità tra episodi, sono state applicate varie normalizzazioni e trasformazioni.

**Normalizzazione temporale**

- Il canale è attivo dal 2009 → forte sbilanciamento nel volume di interazioni nel tempo.
- Soluzione: normalizzazione per anno, in modo da confrontare episodi all'interno dello stesso contesto temporale.



# Gemini data - 5

Creazione della **Metrica Obiettivo Unificata**

Combinazione delle metriche:

- Obiettivo: ottenere una misura unica che rifletta in modo bilanciato la performance del video.
- Metodo: somma ponderata delle tre metriche normalizzate:

Scelte di **ponderazione**:

- Commenti (50%) → indicano interazione attiva e coinvolgimento.
- Like (35%) → rappresentano apprezzamento rapido e diretto.
- Visualizzazioni (15%) → utili ma meno indicative del reale interesse o coinvolgimento.



# Modello base - Regressione senza Grafo

Con il dataset completo, possiamo passare alla valutazione dei modelli.

Obiettivo: **Prevedere la metrica di performance dell'episodio.**

Modelli testati:

- Linear Regression
- Random Forest Regression
- Gradient Boosting Regression

Risultati:

Tra i tre modelli, il **Random Forest** ha ottenuto le migliori performance.



# Modello base - Regressione senza Grafo

La **metrica di riferimento** è **R<sup>2</sup>**, che indica quale percentuale della variabilità della variabile da predire è "spiegata" dal modello.



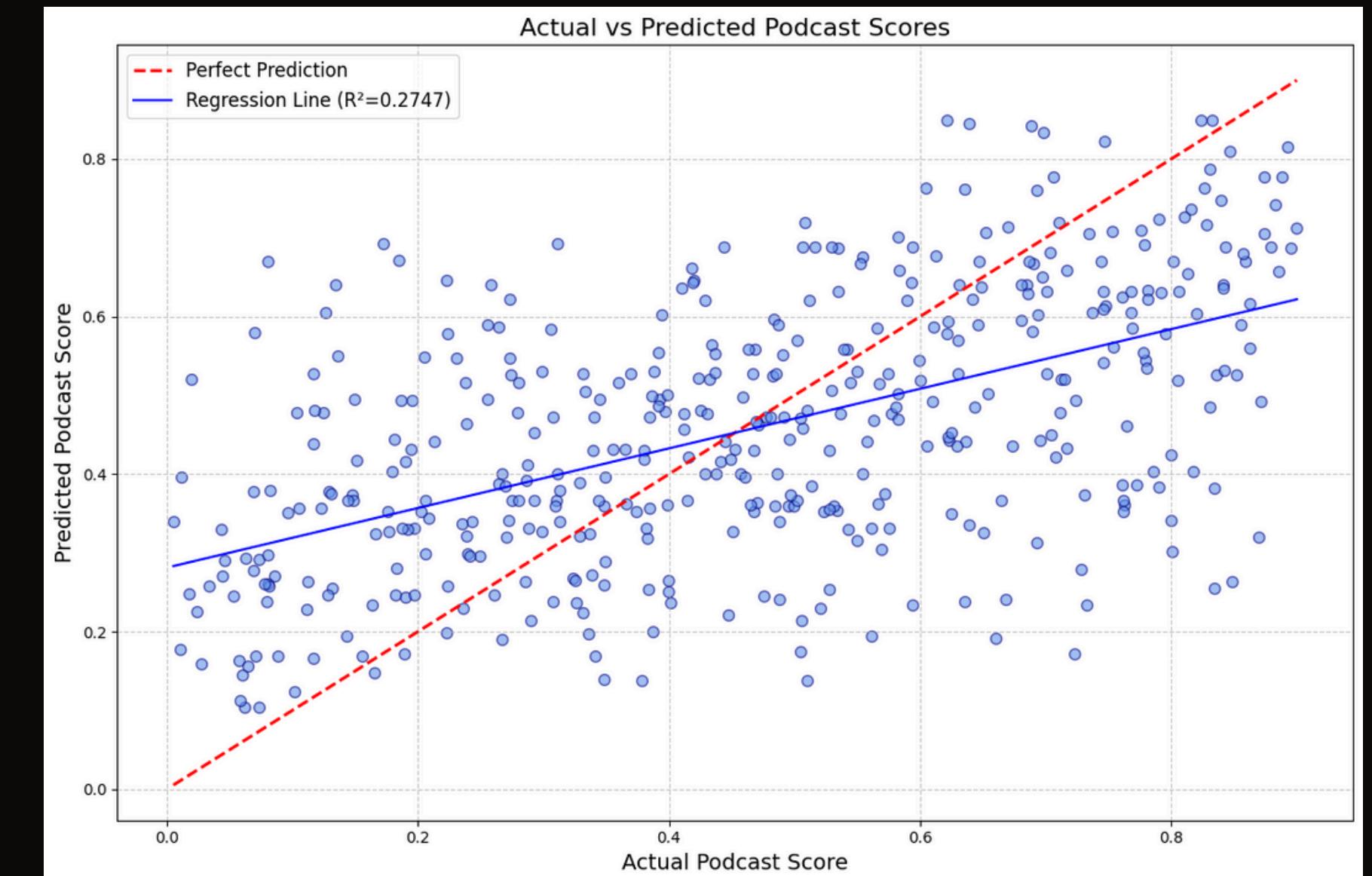


# Modello base - Regressione senza Grafo

Dopo un ottimizzazione dei parametri effettuata tramite il **Grid-search**, il migliore modello ottenuto ha queste performance:

I grafici mostrano che il modello performa molto meglio nel **training set ( $R^2=0,744$ )**, rispetto al **test set ( $R^2=0,274$ )**.

Questo indica che il modello soffre sia di **overfitting** che di **underfitting**.





# Modello avanzato - Regressione con Grafo

Il modello senza grafi ha ottenuto prestazioni molto basse sul test, cerchiamo di migliorare le prestazioni con l'**ausilio dei grafi**.

Utilizziamo l'API di YouTube per ottenere i **nickname dei commentatori** di ogni video e costruiamo un **grafo bipartito** con nodi video e nodi commenter

- Un arco collega un video a un commentatore se quest'ultimo ha commentato quel video.
- Per evitare **data leakage**, il grafo è costruito esclusivamente con i dati del training set (grafo storico).



# Modello avanzato - Regressione con Grafo

A partire dal grafo costruito, ho aggiunto **nuove feature** agli ospiti sfruttando la struttura del grafo. In particolare, per ogni ospite ho calcolato:

- Media del **numero di commentatori** degli ospiti simili
- Media della centralità (**degree centrality**) degli ospiti simili
- Media del **PageRank** degli ospiti simili

Per ospiti simili si intendono quelli che condividono almeno una delle seguenti caratteristiche:

- Nazionalità
- Sesso
- Professione
- Notorietà



# Modello avanzato - Regressione con Grafo

Il nuovo modello, con **feature da grafo**, ha migliorato sensibilmente le prestazioni sul training set, riducendo l'underfitting.

Tuttavia, ha mostrato segni di **overfitting** più marcati.

- Tecniche provate per ridurre l'overfitting:
  - **Grid Search**
  - **Cross Validation**
  - **Feature Selection**

Nessuna ha portato a un miglioramento significativo sul test set.

- **Prestazioni finali:**
  - $R^2$  training set: 0.928
  - $R^2$  test set: 0.137



# Modello avanzato - Regressione con Grafo

Per indagare la **scarsa capacità di generalizzazione** del modello, ho costruito un **grafo degli ospiti**.

- Ogni nodo rappresenta un ospite, e due ospiti sono collegati se hanno **commentatori in comune**.
  - → Il peso degli archi corrisponde al numero di commentatori condivisi.

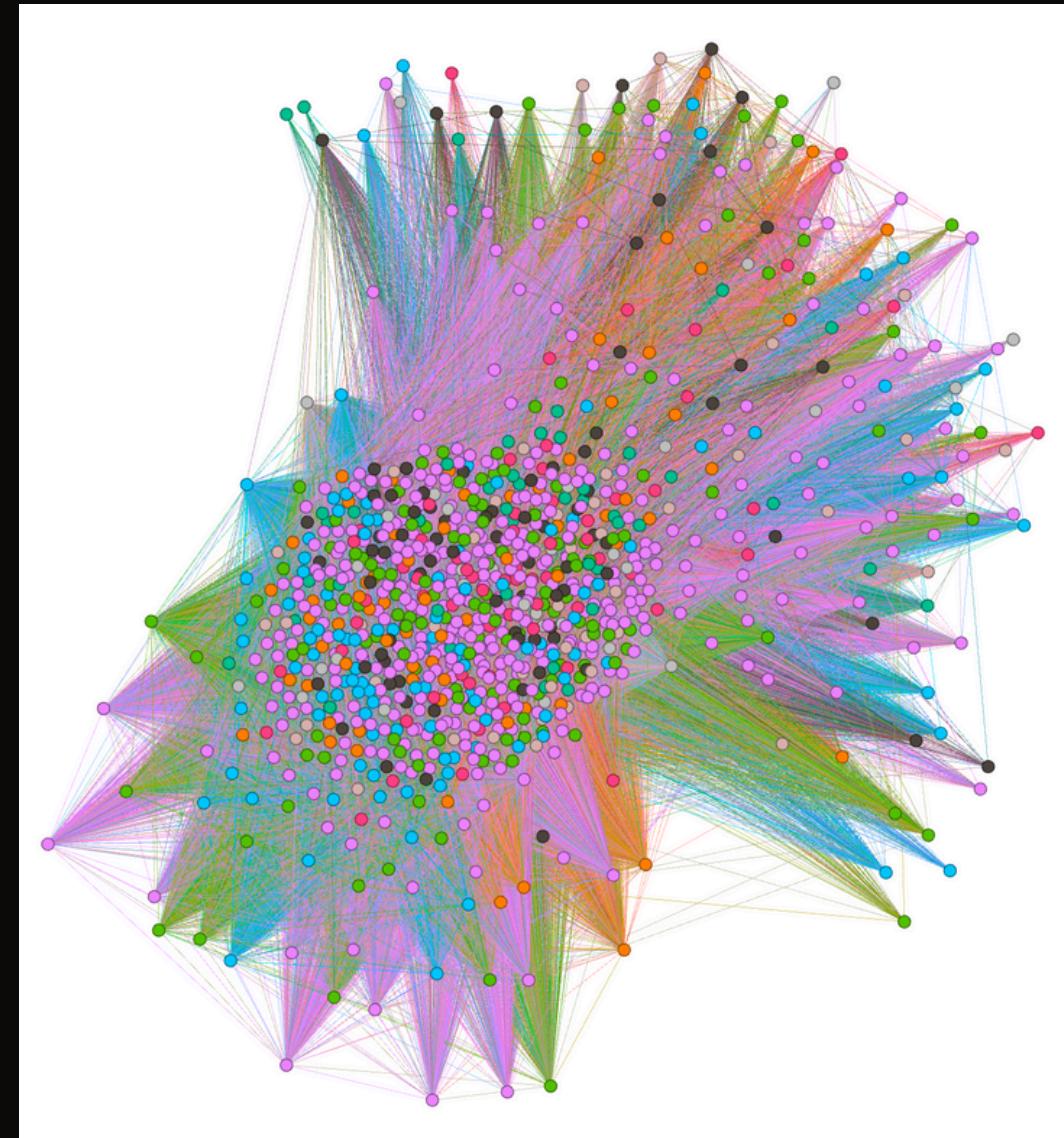
L'obiettivo: verificare se esistono **community** coerenti con caratteristiche come:

- Nazionalità
- Professione
- Notorietà
- Topic del video

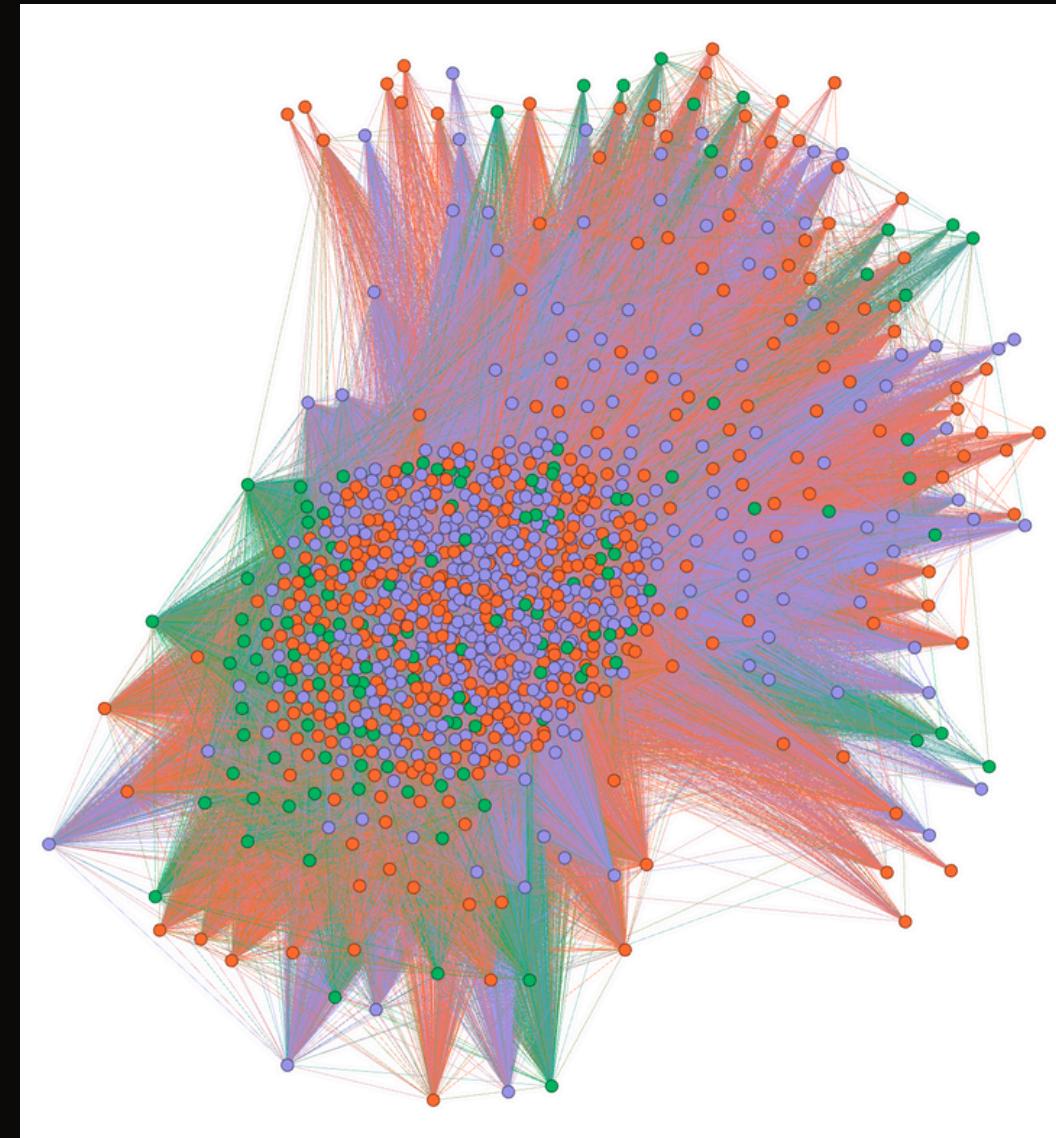


# Modello avanzato - Regressione con Grafo

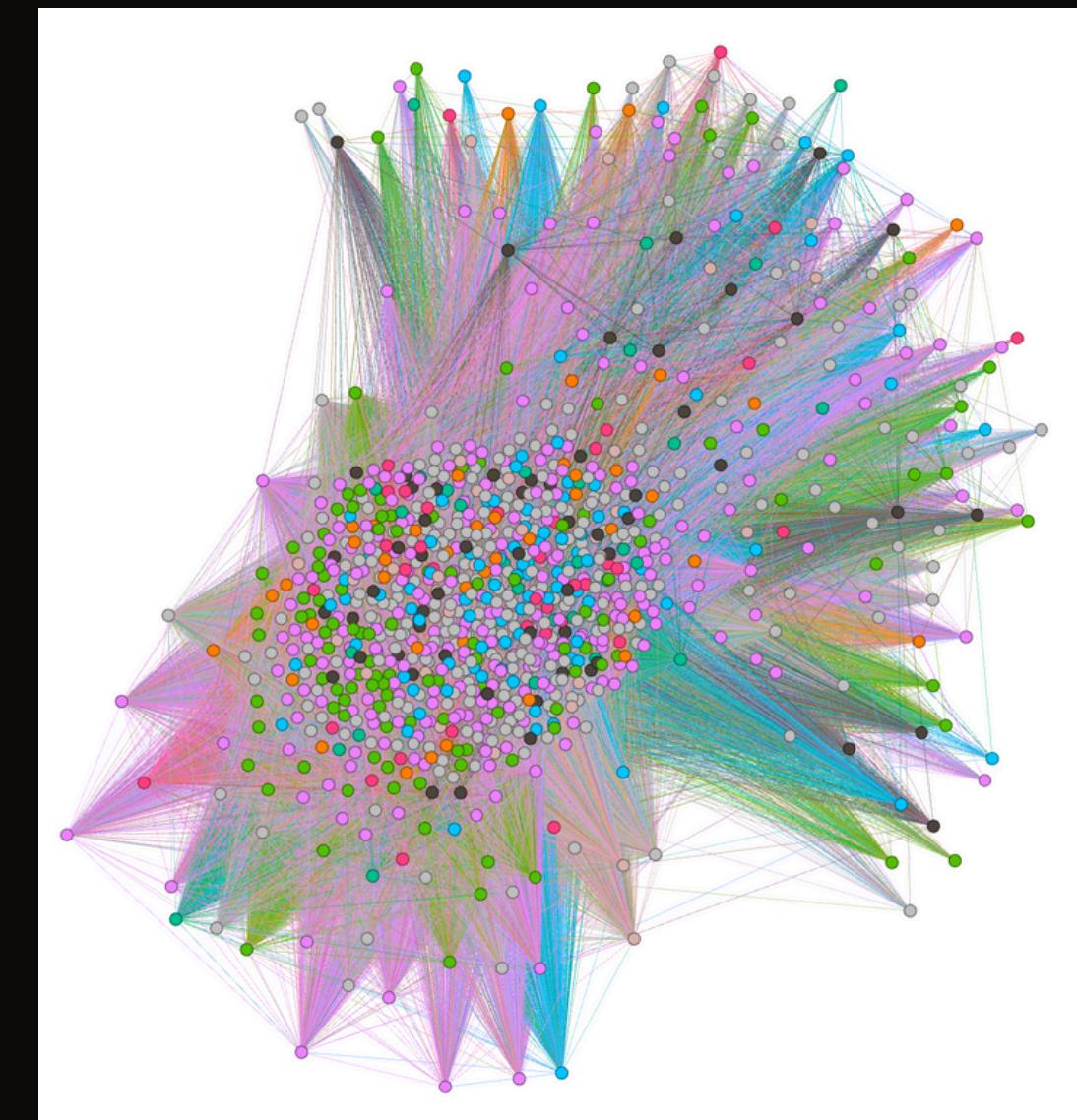
Professione:



Notorietà:



Topic video:





# Modello avanzato - Regressione con Grafo

Dai grafi generati **non emergono divisioni chiare in community** basate sulle caratteristiche analizzate.

Tutti gli ospiti tendono a formare un **unico grande cluster**.

- Questo indica che la struttura relazionale tra gli ospiti (in base ai commentatori) non riflette le loro caratteristiche semantiche.

Conclusione: le feature derivate dal grafo non aggiungono una separazione informativa utile alla generalizzazione del modello.



# Conclusioni e sviluppi futuri

Il modello attuale non riesce a generalizzare, probabilmente a causa della **scarsa diversità strutturale** nel grafo e della mancanza di segnali informativi forti.

Possibili **miglioramenti futuri**:

1. Scegliere caratteristiche degli ospiti diverse e più utili al task;
2. Fare data augmentation per avere un dataset più ampio e bilanciato;
3. Sfruttare il sentimento dei commenti.