L’endometriosi e il cibo

Sara Galatro – Febbraio 2023

Abstract

L'obiettivo di questo progetto è analizzare la relazione tra varie persone che soffrono di endometriosi e diciannove cibi selezionati, così da impostare un **sistema di raccomandazione** tra pazienti.

I dati, raccolti da me stessa tramite un form Google, sono stati poi elaborati con due metodi diversi: il **K-Nearest Neighbour standardizzato** e la **Singular Vector Decomposition**.

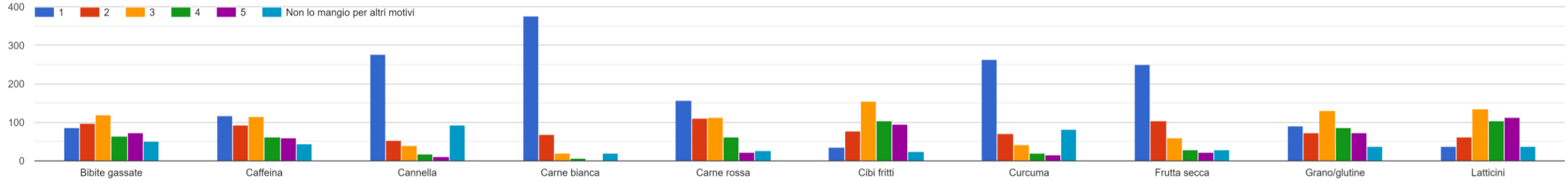
I risultati ci permettono di delineare alcuni dettagli interessanti e confermano che, con le giuste modifiche, un sistema del genere sarebbe davvero molto di aiuto a coloro che ne necessitano.

Motivazione

L'**endometriosi** è una malattia cronica che debilita fortemente le persone che ne soffrono, sia fisicamente che psicologicamente. Al momento, **non vi è cura**, ma solo delle terapie volte a limitare i sintomi di tale patologia, come ad esempio un determinato tipo di alimentazione.

Questa dieta è stata studiata da un punto di vista prettamente chimico-biologico, ma coloro che soffrono di endometriosi hanno sempre fatto notare come per loro queste linee guida non solo **non funzionassero**, ma addirittura le facessero star peggio.

Si vuole dunque analizzare l'alimentazione confrontando le reazioni delle pazienti ai vari cibi, **spostando così il focus dall'alimento alla persona**, e costruire un sistema che riconosca esperienze simili e che sia in grado di consigliare nuovi cibi in base ad esse.

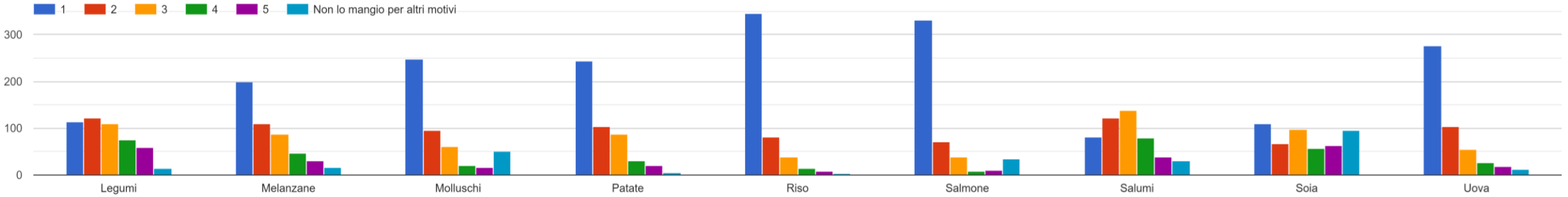


Dataset

I dati sono stati raccolti tramite un form Google e pubblicato su diversi gruppi social creati come supporto da e per le persone che soffrono di endometriosi.

In **492** hanno risposto al sondaggio, in cui veniva loro chiesto:

- l'età;
- da quanto soffrano di endometriosi;
- di dare una valutazione da 1 a 5 ai diciannove cibi selezionati, dove 1 indica che il cibo non fa male e 5 indica che fa molto male (secondo la propria esperienza).



Preparazione e pulizia dei dati

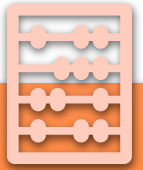
Prima formattazione

- I valori della colonna «*Da quanto tempo soffre di endometriosi?*» sono stati codificati in **numeri**
- Le risposte «*Non lo mangio per altri motivi*» sono state **codificate come 0** per escluderle dall'analisi
- Tutti i voti sono stati convertiti per praticità in **interi**

Analisi preliminare

- Da un'analisi preliminare dei dati ho dedotto che
  a) lo spazio dei voti è eccessivamente piccolo
  b) i voti non dipendono né dall'età né dal periodo di diagnosi
- Per distinguere meglio i voti, **moltiplichiamo** i dati per cinque
- **Eliminiamo** le colonne «*età*» e «*Da quanto soffre di endometriosi?*»

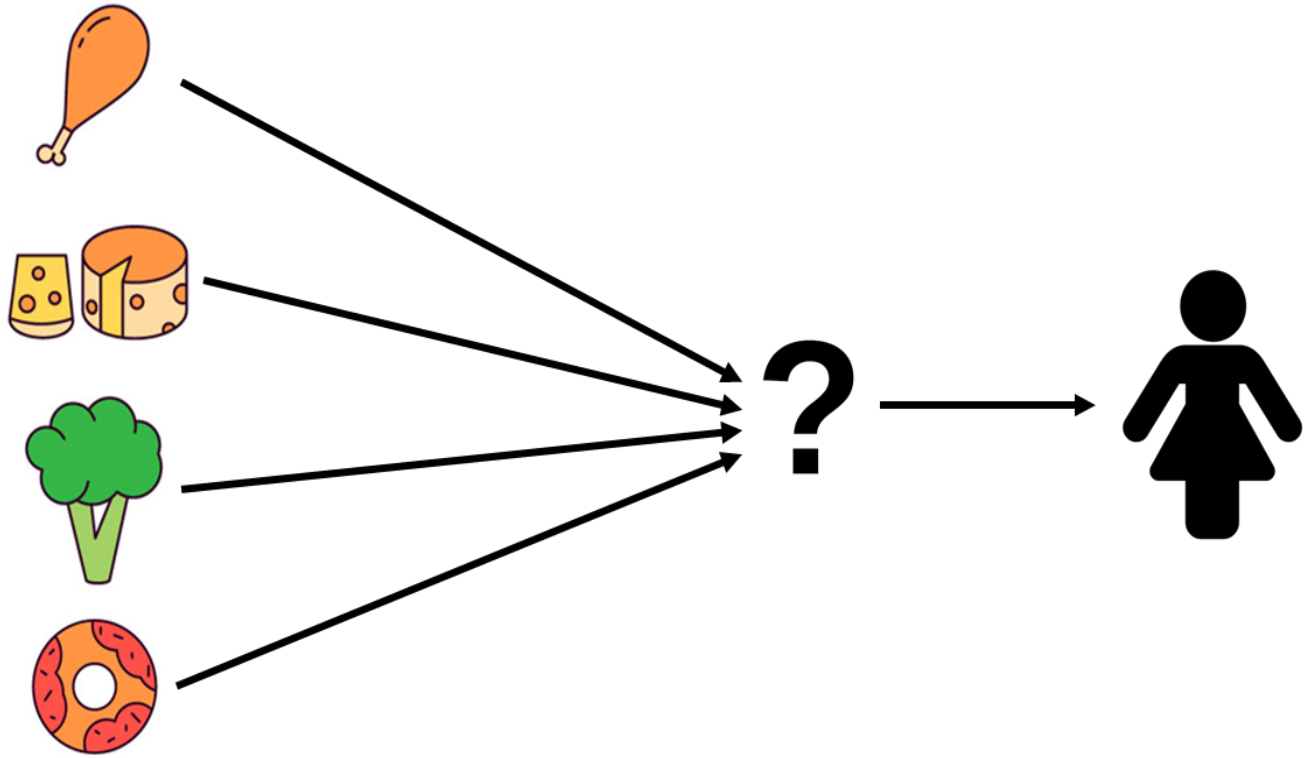
Formattazione Surprise



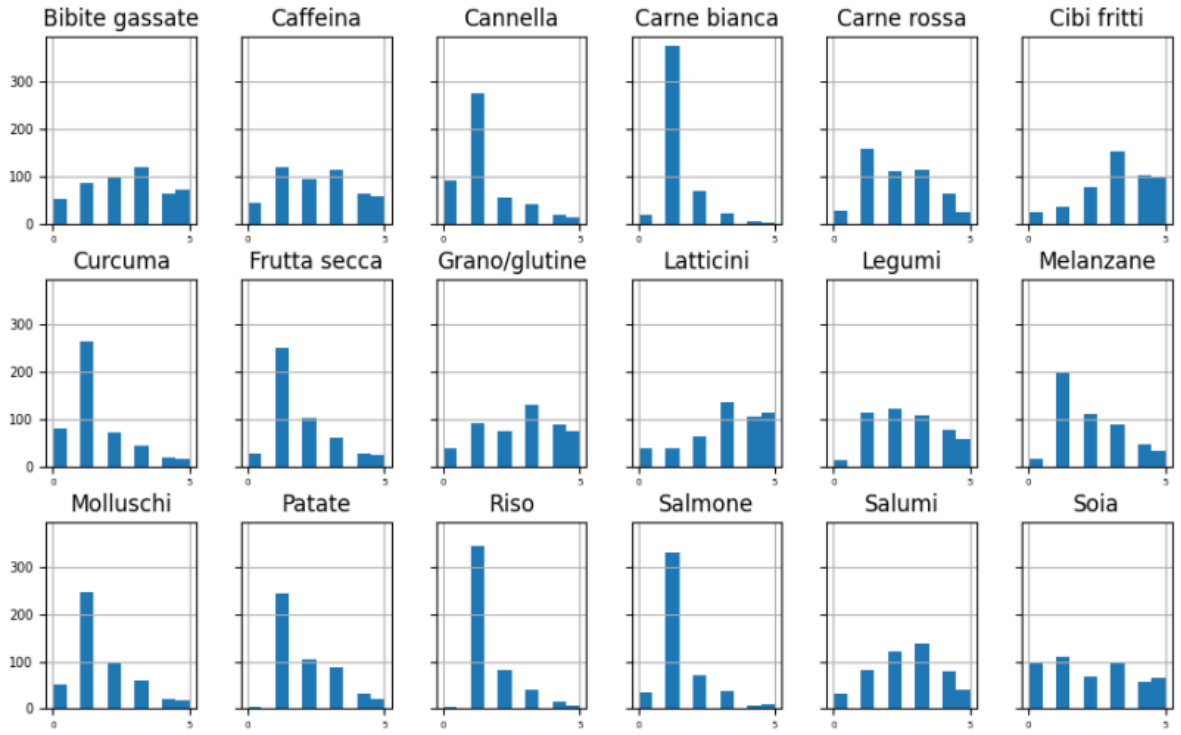
- Il dataframe ausiliario deve avere, per ogni riga, tre entrate nell'ordine
  [user_id][food_id][rating]
- Riformattiamo i dati per avere questo layout, con l'accortezza di escludere i valori nulli
- Costruito l'oggetto «Dataset», possiamo dividere i nostri dati in **training set** e **test set**, con una proporzione di 80%-20%

Obiettivo della ricerca

Una volta verificato che i dati raccolti non separano così nettamente i cibi come la dieta promette, l'obiettivo di questa ricerca è capire se un **sistema di raccomandazione** possa essere uno strumento utile per pazienti ed esperti nella creazione di regimi alimentari personalizzati e mirati.



Analisi preliminare dei dati



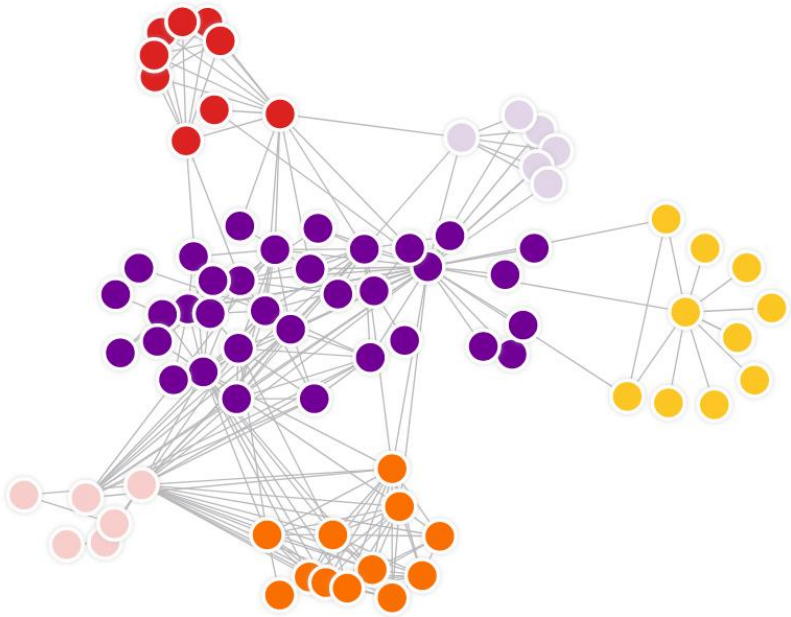
Oltre a verificare l'esistenza o l'assenza di una correlazione, un primo importante risultato è dato da una semplice rappresentazione dei dati raccolti.

Ricordando che con il valore «0» si indica i voti non pertinenti alla nostra indagine, dagli istogrammi dei singoli cibi si evince che, ad eccezione di pochi casi, che **non c'è una distinzione netta tra i cibi** e che quindi coloro affetti da endometriosi hanno davvero delle linee guida eccessivamente generiche a loro disposizione.

Tentativo 1: K-Nearest Neighbour

Il primo tentativo è stato quello di utilizzare il modello più intuitivo: dato un utente con delle valutazioni sconosciute, cerchiamo gli utenti più **simili** in base alle valutazioni note. Una volta posizionato l'utente nello spazio dei cibi, restituiamo per i suoi valori mancanti la media dei suoi vicini per quei determinati cibi:

$$\hat{r}_{ui} = \mu_u + \sigma_u \frac{\sum_{v \in N_i^k(u)} \mathrm{sim}(u,v) \cdot (r_{vi} - \mu_v)/\sigma_v}{\sum_{v \in N_i^k(u)} \mathrm{sim}(u,v)}$$



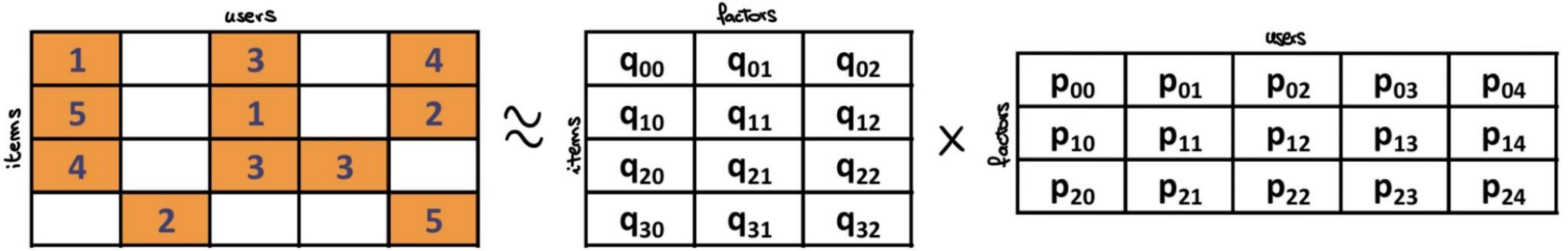
Inoltre, visto lo spazio ridotto in cui sono concentrati gli utenti, si è preferito **normalizzare** le valutazioni, così da dare maggior peso alle variabili che più di tutte identificano la posizione del paziente.

Tentativo 2: SVD

Il secondo modello è meno intuitivo ma più efficace, sia in termini di risultati che di tempo di esecuzione: l'algoritmo SVD si pone di fattorizzare una matrice di input $\boldsymbol{R}$ come

$$\boldsymbol{R} = \boldsymbol{U\Sigma V^T} \equiv \boldsymbol{QP^T}$$

Graficamente:



Tentativo 2: SVD

Nel caso dei sistemi di raccomandazione, la matrice $\boldsymbol{R}$ è incompleta: l'algoritmo si pone dunque di ricostruire le due matrici $\boldsymbol{Q}$ e $\boldsymbol{P}$ minimizzando la funzione costo regolarizzata

$$\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda(\|q_i\|^2 + \|p_i\|^2)$$

tramite una semplice discesa stocastica del gradiente, ossia aggiornando per ogni esempio i parametri (ossia le entrate delle matrici) secondo la regola

$$p_u \leftarrow p_u + \eta(\epsilon_{ui} \cdot q_i - \lambda p_u)$$

$$q_i \leftarrow q_i + \eta(\epsilon_{ui} \cdot p_u - \lambda q_i)$$

Metriche di accuratezza

Per entrambi i modelli sono state utilizzate le seguenti metriche di accuratezza:

- il **Mean Absolute Error**, definito come

$$mae = \sum_{k=1}^{m} |\hat{r}_k - r_k|$$

- la **Fraction of Concordant Pairs**, definita come

$$fcp = \frac{n_c}{n_c + n_d} \qquad \text{dove} \qquad n_c = \sum_u n_c^u$$

Con coppie concordanti si intende le coppie di valutazioni le cui stime rispettano l'ordine dei voti originali, ossia

$$n_u^c = |\{(i,j) : \hat{r}_{ui} < \hat{r}_{uj} \text{ and } r_{ui} < r_{uj}\}|$$

Risultati

Per valutare l'accuratezza complessiva e interpretare praticamente i risultati del nostro sistema, si è concentrata l'attenzione sugli «**errori gravi**»: nell'ottica di un'implementazione futura e interfacciandosi con veri utenti, il sistema dovrebbe essere in grado di identificare tre tipologie di cibo in base ai voti e consigliarli/sconsigliarli in base a tale classificazione.

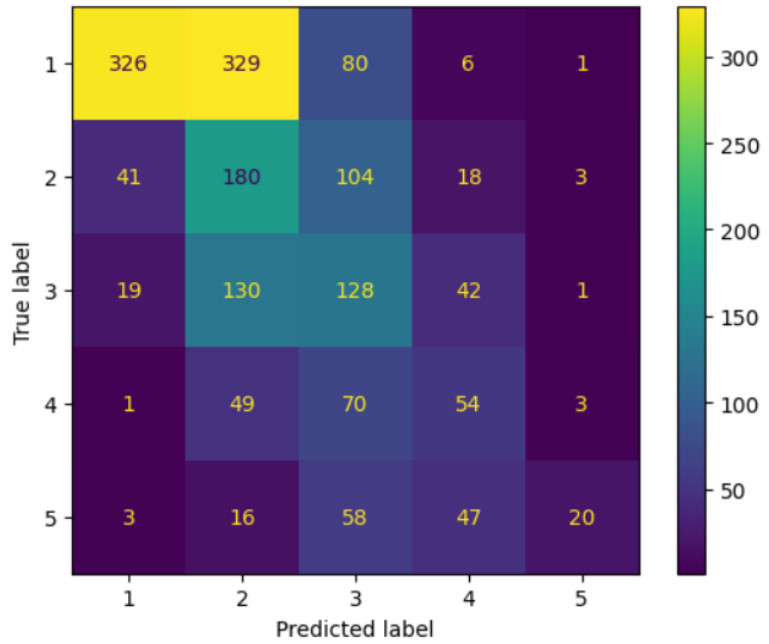
Voto: 1 o 2 Cibo sicuro	Voto: 3 Da consumare limitatamente	Voto: 4 o 5 Cibo non sicuro

Pertanto, non ci interessa che il sistema sia in grado di predire correttamente i voti dei pazienti, ma solo che non sbagli la pericolosità del cibo analizzato.

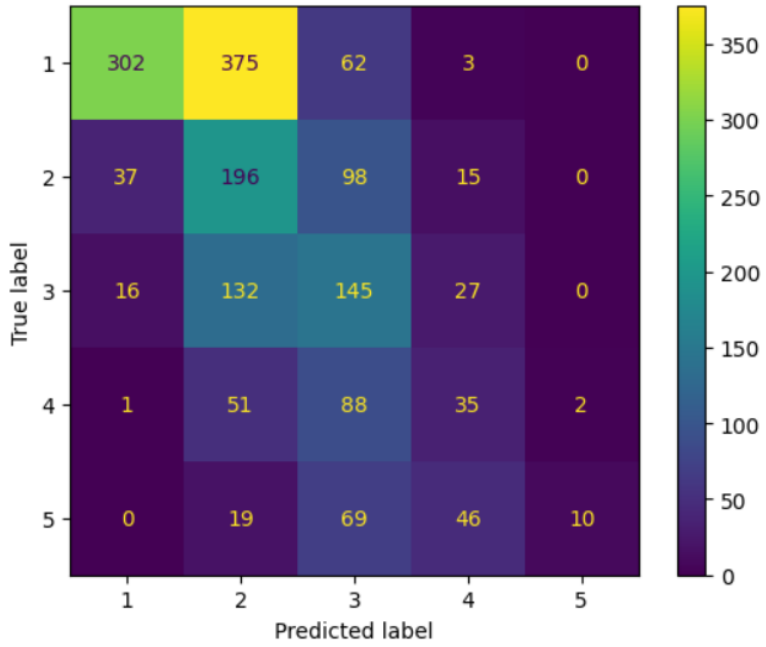
Con questa premessa, il metodo K-NN ha effettuato una percentuale di errori gravi pari al 14.75%, mentre il metodo SVD ha ridotto tale percentuale a 13.65%.

Risultati

Riportiamo per completezza anche le matrici di confusione dei due modelli:



K-NN

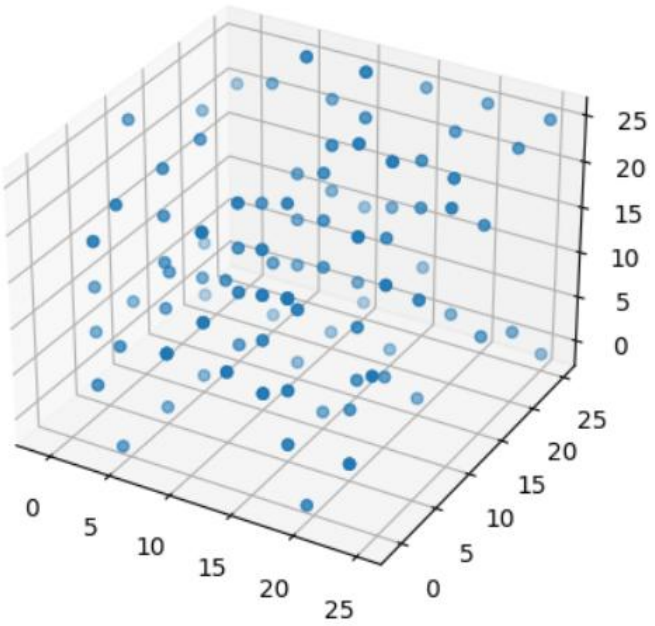


SVD

Limitazioni

È importante notare che una delle difficoltà maggiori sia la **dimensione del nostro spazio cibi** (dim = 19) rispetto alle valutazioni (da 1 a 5): questa sproporzione fa sì che tutti gli utenti siano condensati in un volume ridotto e distribuiti abbastanza uniformemente. Ciò fa sì che molte informazioni sulla correlazione dei dati siano perse e peggiora la separazione dei pazienti nello spazio.

Inoltre, alcune categorie potrebbero presentare più **rumore** di altre raccogliendo diverse tipologie di cibo.



Conclusioni e lavori futuri

Fin dall'analisi preliminare è emerso come ci sia bisogno di un **sistema ausiliario** per progettare meglio le terapie alimentari di coloro che soffrono di endometriosi.

Il sistema di raccomandazione qui proposto dà **risultati incoraggianti**: nonostante la quantità e la qualità limitata dei dati, gli algoritmi sono in grado di riconoscere abbastanza similitudini tra gli utenti da evitare di commettere troppi errori gravi.

È dunque logico pensare che migliorando i dati e analizzandone di più il sistema sarà in grado di aumentare la sua accuratezza e la sua sicurezza.

In un lavoro futuro e più completo si potrebbero anche aumentare le classi di cibo, includendo vari sotto-casi, così da ridurre il rumore delle singole valutazioni.

Inoltre, dato l'obiettivo di tale sistema, potrebbe essere utile provare a implementare un sistema che si basi completamente sulla misura *fcp,* come ad esempio quello basato sull'algoritmo **OrdRec** di Koren et Sill, tenendo così maggior conto della classifica dei cibi per utente piuttosto che del valore numerico di tali valutazioni.

Riconoscimenti

Ringrazio tutte coloro che hanno accettato di partecipare al sondaggio e di aver condiviso le loro esperienze con me, permettendomi di raccogliere così tanti dati in così poco tempo. Si ringrazia inoltre il gruppo «**Endometriosi – Community**» per avermi permesso di usare il loro gruppo per far girare il sondaggio.

Ringrazio infine il mio amico e collega Pietro Cestola, che ha accettato di sentire la mia presentazione e di darmi la sua opinione sul mio lavoro.

Riferimenti bibliografici

- Capitolo 9 del libro "*Mining of Massive Datasets*", Leskovec et al, 2014
- Documentazione della libreria Surprise
- "*Collaborative Filtering on Ordinal User Feedback*", Yehuda Koren e Joseph Sill, 2011