SPOTIFY RECOMMENDATION SYSTEM

HOMEWORK II – STATISTICA COMPUTAZIONALE
APPELLO DEL 27/04/2023
SARA NAVA (870885), GIULIA SARESINI (864967), NICOLA PERANI (864755)

SPOTIFY

Spotify è un **servizio di streaming musicale** disponibile via app e sito web che dà accesso a milioni di brani, podcast e video di artisti di tutto il mondo. Una delle potenzialità di questo servizio è il suo **recommendation system**: sulla base degli ascolti e delle preferenze dell'utente, infatti, è in grado di personalizzare la sua esperienza facendogli scoprire ogni giorno nuovi brani e contenuti che possano interessargli.

OBIETTIVO DEL PROGETTO

L'obiettivo di questo progetto è quello di cercare di stimare, sulla base delle conoscenze finora apprese, un **modello in grado di classificare i brani offerti dalla piattaforma in due classi** (*Dislike*, *Like*) **a partire dalle preferenze dell'utente**, per poter prevedere quali contenuti musicali possano essere di suo interesse.

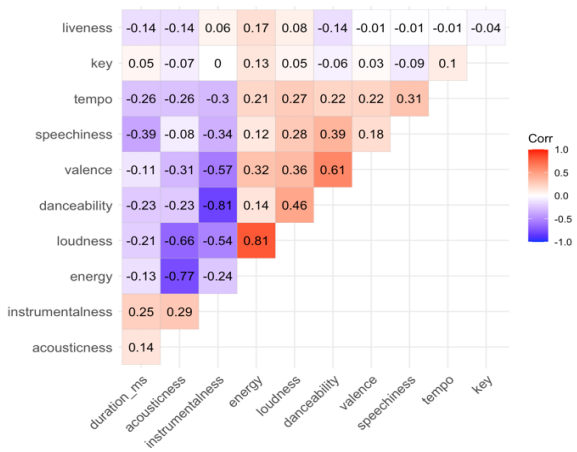
DESCRIZIONE DEL DATASET

Il dataset è stato costruito a partire dall'interrogazione dell'**API ufficiale di Spotify** da parte di un utente, il quale ha scaricato **13** metriche (**variabili** quantitative e qualitative) per **195 brani** da lui scelti, classificati poi nelle **classi *"Dislike"* e *"Like"*** (per maggiori informazioni sul campionamento si rimanda alla pagina web GitHub del creator) sulla base dei suoi gusti musicali.

ANALISI ESPLORATIVA

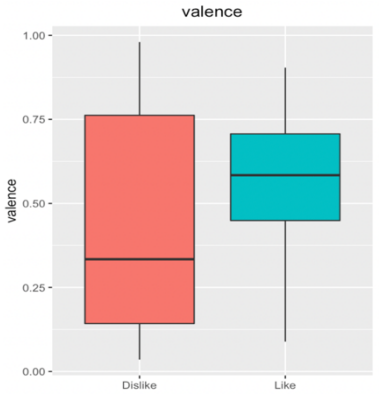
Analisi della correlazione.
Il dataset è caratterizzato da **variabili non fortemente correlate tra loro**, fatta eccezione per la variabile *instrumentalness* che risulta fortemente correlata negativamente con almeno la metà delle variabili. Una bassa correlazione tra variabili potrebbe impattare positivamente sulla stima dei modelli di clustering, in quanto si riduce il rischio di multicollinearità (e quindi di instabilità del modello stimato).

Distribuzione delle due classi.
In linea generale, i **gruppi** sembrerebbero avere **caratteristiche differenti in termini di media e varianza**, aspetto che potrebbe portare a risultati

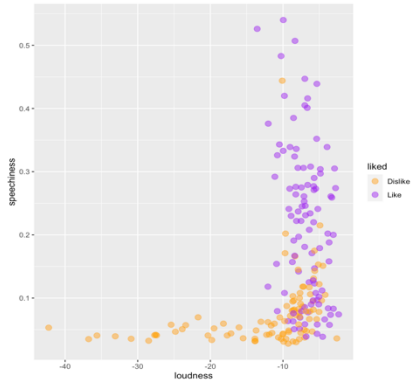


1. Grafico delle correlazioni

migliori nella fase di stima dei modello di clustering.
Tuttavia, alcune variabili presentano un elevato numero di osservazioni anomale e/o una distribuzione non gaussiana. Si è tentato quindi di applicare una trasformazione adeguata alle variabili maggiormente problematiche, ottenendo però diversi insuccessi: in alcuni casi, la distribuzione non perdeva le osservazioni anomale, mentre in altri la distribuzione si uniformava eccessivamente rispetto alle classi.



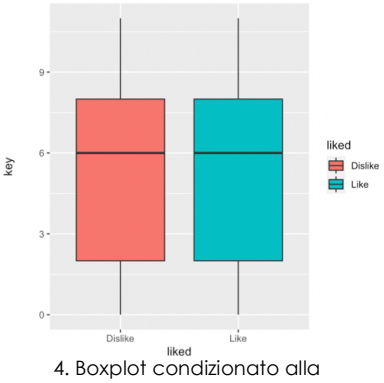
2. Boxplot condizionato alla classe della variabile *valence*.



3. Grafico a dispersione rispetto alle variabili *loudness* e *speechiness*.

Per questo motivo, i modelli sono stati stimati a partire dai dati originali, tenendo conto della presenza di eventuali anomalie in fase di interpretazione dei risultati.
Osservando, inoltre, i grafici di dispersione, si può notare che in molte coppie, i cluster erano discretamente distinti, fatta qualche eccezione per la presenza di una **zona di sovrapposizione**; in altre coppie, invece, la distinzione tra i due gruppi era meno evidente, mostrando quindi una distribuzione quasi casuale. Entrambi i grafici illustrano come la distribuzione effettivamente sia distinta nelle due classi (nei casi più significativi), ma di come a volte si crei un'area di interferenza.

FEATURE SELECTION

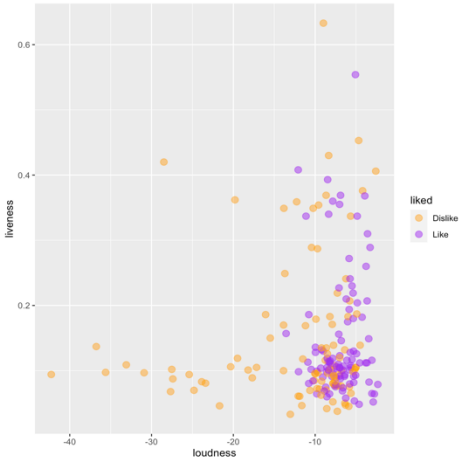


4. Boxplot condizionato alla classe della

A causa della natura dei modelli stimati è stato necessario sin dal principio **escludere** dall'analisi le 2 **variabili qualitative** (di tipo categorico) presenti nel dataset (*mode* e *time_signature*). Anche la variabile *key* è stata esclusa: in fase di analisi esplorativa è emersa una distribuzione analoga nelle due classi (*fig. 4*) e sarebbe stata ridondante a livello di clustering.
La rimozione di queste, probabilmente, non impatterà significativamente sul clustering poiché si tratta di variabili legate al suono in generale, dunque ipoteticamente indipendenti dalle preferenze musicali di un utente (si rimanda alla documentazione dell'API per una descrizione più dettagliata delle variabili).

È stata poi **rimossa** la variabile *instrumentalness* a causa **dell'andamento eccessivamente** sbilanciato e **anomalo per la modalità "*Like*"** e una **correlazione** più **alta** della media assoluta con almeno la metà delle variabili del dataset, come emerso in fase di analisi esplorativa.

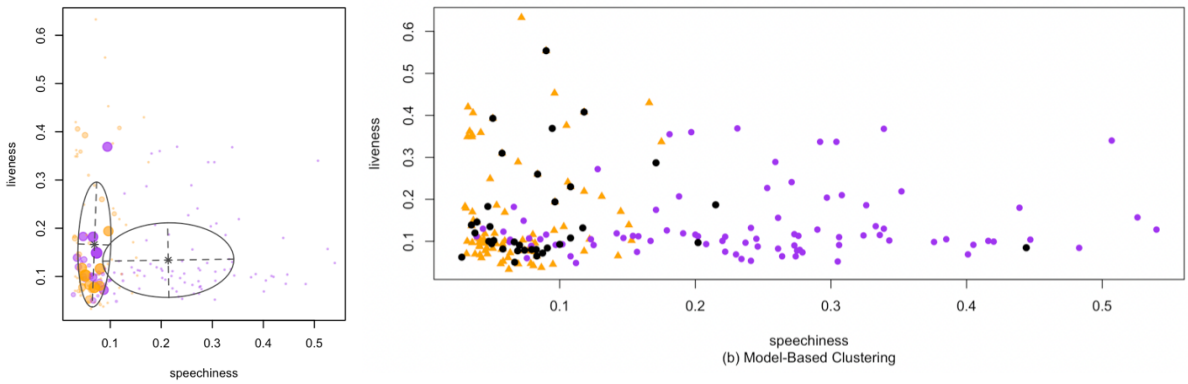
Tramite **PCA** sono state individuate le **3 variabili** che incidono maggiormente rispettivamente sulle prime tre componenti principali (che spiegano il 67% della varianza) per capire se effettuare un'ulteriore riduzione delle dimensionalità. Osservando però gli scatterplot delle combinazioni di coppie formate dalle variabili selezionate, si può notare che non tutte distinguono bene i cluster (es. fig. 5) ma sembrerebbero mostrare una distribuzione casuale. Per questo motivo, si procede tenendo in considerazione tutte le variabili rimaste, quindi un dataset di 195 osservazioni e 10 variabili, di cui una è quella di classificazione.



5. Scatterplot delle variabili *loudness* e *liveness*.

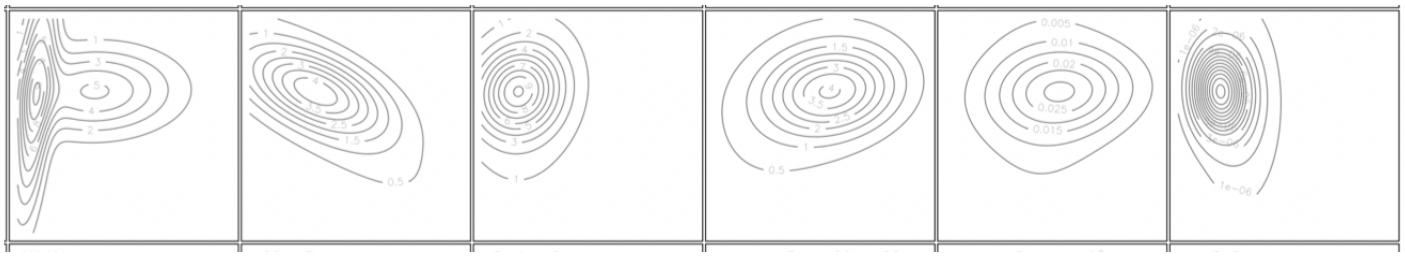
MODEL BASED CLUSTERING

Tramite tecniche di clustering model – based si è cercato di identificare pattern di similarità nei dati, esplorando la struttura sottostante per la comprensione degli stessi. Per farlo si è proceduto applicando tali tecniche al fine di stimare i parametri della mistura di **2 componenti** (numero fissato a priori vista la suddivisione originale dei dati in due gruppi) che meglio rappresenta la distribuzione dei dati, e tramite la minimizzazione dell'**ICL** e del **BIC**, è stato selezionato come miglior **modello per la struttura di variabilità della mistura** un **VVE**: sono stati quindi individuati **due cluster di volume e forma variabile** e con **medesimo orientamento**.
Il modello riporta performance discrete ma non ottime, poiché applicando il modello ai dati circa il **18%** delle osservazioni risulterebbe **mal classificato**: si osserva però che queste rientrano nella zona di sovrapposizione individuata in fase di esplorazione dei dati (fig. 7), il che giustifica un'incertezza maggiore (fig. 6) nell'assegnazione e la possibilità di errori.



Da sinistra:
6. Grafico dell'incertezza associata all'assegnazione ai cluster per le variabili *speechiness* e *liveness*.
7. Scatterplot che mostra (in nero) le unità mal classificate, coi valori delle variabili *speechiness* e *liveness*.

Illustrando la struttura di clustering mediante le curve di livello (fig. 8), si nota la sovrapposizione dei gruppi per la maggior parte di coppie di variabili, e la presenza di **curve non sempre regolari** (ellittiche o circolari). Questo potrebbe suggerire, in fase di stima dei modelli di classificazione, un **miglior funzionamento di modelli** che **non prevedono una distribuzione gaussiana** (multidimensionale) associata a ciascun cluster.



8. Alcuni esempi di curve di livello per coppie di variabili differenti: è evidente come non sempre vi sia circolarità e regolarità delle curve, ma soprattutto la fusione dei due cluster.

CLASSIFICATION

Sono state applicate due tecniche di classificazione di tipo probabilistico che utilizzano un approccio generativo, ovvero **EDDA** (*Eigenvalue Decomposition Discriminant Analysis*) e **MDA** (*Mixture Discriminant Analysis*). La tecnica che ha riportato una più accurata capacità di classificazione dei dati è stata quest'ultima, **MDA**.

Nell'applicazione delle tecniche di classificazione si è deciso di non testare tutti i possibili modelli per la struttura di variabilità della mistura (sia per MDA che per EDDA) ma, piuttosto, di tenere in considerazione solo i corrispondenti analoghi alla struttura individuata nei dati a disposizione in fase di applicazione delle tecniche di clustering, ovvero ***VVE***.

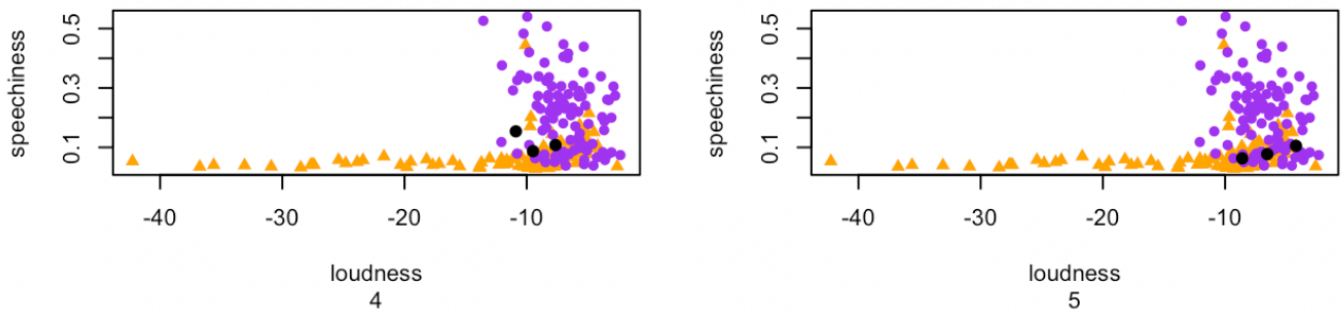
Si è quindi specificato per la tecnica MDA il modello ***VVE*** e per **EDDA** il modello ***Gaussian_p_Lk_D_Ak_D***.

Per entrambi i metodi, date le dimensioni ridotte del dataset, si è poi deciso di applicare la tecnica di ***Cross-Validation,*** suddividendo il dataset in cinque campioni aleatori di 39 unità ciascuno.
In questo modo si possono valutare le prestazioni del modello su diverse combinazioni di training e test set, migliorando la validità delle analisi.
Questo passaggio, che potrebbe sembrare poco utile dato il funzionamento delle funzioni *mixmodLearn* e *MclustDA* che applicano già una *Cross-Validation,* si è reso necessario data la forte dipendenza trovata tra il *seme* settato nella generazione dei campioni casuali e l'*error rate* ottenuto (che nella maggior parte dei casi influiva non solo su quest'ultimo, ma anche sul modello che risultava essere il migliore).
Non si può garantire in modo assoluto la correttezza di questi risultati, ma ci si aspetta che il procedimento applicato li renda almeno relativamente affidabili.
In sostanza, si ottiene un range di variazione per l'*error rate* che, a differenza del singolo valore, risulta essere più affidabile, poiché meno influenzato dalla possibile estrazione di un training set "particolare".

EDDA

Il modello per la struttura di varianza della mistura di distribuzioni gaussiane multivariate che si è deciso di testare è il ***Gaussian_p_Lk_D_Ak***.
Questo modello riporta buoni risultati nella stima di misture aventi le stesse *mixing probabilities*, le cui component density sono associate a cluster di forma e volume differenti e stesso orientamento.

Si è calcolato l'*error rate* sulla prima combinazione di training e test set e si è ottenuto un valore pari a **0.1538462**. Il risultato è buono, soprattutto perché è importante considerare che la natura dei dati e delle variabili a disposizione rende difficile distinguere in modo netto le due classi.
Al fine di confermare la validità dei risultati ottenuti è stato valutato il modello su diverse combinazioni di training e test set ottenendo un range di variazione dell'*error-rate* pari a **[0.07, 0.21]**.



9. Il grafico riportato mostra, in nero, le osservazioni mal classificate per 2 campioni di test set (in particolare i campioni 4 e 5) sulle variabili *loudness* e *speechiness*. In generale si può quindi dire che l'errore della classificazione si concentra nelle zone di sovrapposizione delle osservazioni.

Si noti che la maggior parte dei dati mal classificati, per ogni combinazione di training e test, si trova all'incirca nella zona di sovrapposizione individuata in fase di analisi esplorativa (nel paragrafo dedicato alle conclusioni verrà fornita una possibile spiegazione a questo fenomeno e un metodo di raccolta dati che avrebbe potuto ovviare a questo problema).

MDA

Per la Mixture Discriminant Analysis si è scelto di utilizzare il modello ***VVE*** che, proprio come il modello *Gaussian_p_Lk_D_Ak_D* utilizzato nella tecnica di classificazione EDDA, per effettuare

appunto una classificazione su dati distribuiti in cluster di forma e volume differenti e uguale orientamento.

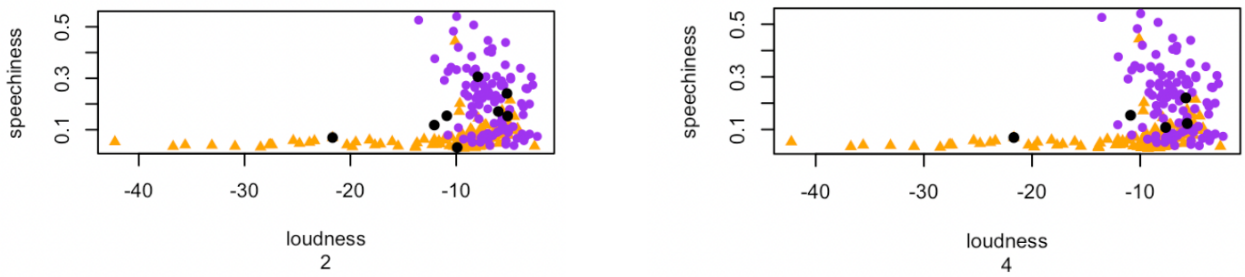
Nella seguente tabella sono mostrati i valori di *error rate* ottenuti su ciascun campione applicando le due tecniche di classificazione:

CAMPIONE	ERROR RATE – EDDA	ERROR RATE – MDA
1	0.1538462	0.1282051
2	0.2051282	0.2051282
3	0.1025641	0.07692308
4	0.07692308	0.1282051
5	0.07692308	0.05128205

10. Tabella che consente di confrontare gli error rates ottenuti applicando le due tecniche di classificazione per ogni campione generato.

Osservando quindi i risultati mostrati in tabella si può affermare che le **performance di MDA** sono **migliori rispetto a quelle di EDDA**, ma con una differenza minore rispetto a ciò che si sarebbe potuto affermare se ci si fosse fermati al primo valore dell'*error rate*.
Nota: non si può avere la certezza che generando i campioni impostando un seed diverso, i risultati sarebbero rimasti gli stessi.

Anche in questo caso, come nel modello precedente, le osservazioni classificate non correttamente si trovano nella zona di sovrapposizione tra le due classi.



11. Il grafico riportato mostra, in nero, le osservazioni mal classificate per 2 campioni di test set (in particolare i campioni 2 e 4) sulle variabili *loudness* e *speechiness*. In generale si può quindi dire che l'errore della classificazione si concentra nelle zone di sovrapposizione delle osservazioni.

CONCLUSIONI

Sia per il clustering che per la classificazione il problema che si è riscontrato maggiormente riguarda la **sovrapposizione delle osservazioni appartenenti a classi differenti**.

Questo è causato dal fatto che, come mostrato in fase di analisi esplorativa, la classe "*Dislike*" presenta una varianza maggiore rispetto alla classe "Like", dunque, una maggiore ampiezza dell'intervallo di valori assunti che spesso includeva quello della classe delle canzoni apprezzate. Di conseguenza la maggior incertezza nell'allocazione è associata proprio alle osservazioni della classe "*Dislike*".

Si può quindi ipotizzare che, nel caso dell'utente in esame, i gusti musicali siano ben definiti: i suoi ascolti si limitano ad un insieme di brani che probabilmente riconducono ad uno stesso genere musicale, al contrario di quelli da lui non apprezzati che potrebbero far parte di un insieme di più generi.

La sovrapposizione potrebbe essere dovuta al fatto che i gusti musicali, di solito, più che basarsi sul tipo di rumore, la durata della canzone e altre caratteristiche legate alle variabili presenti nel

dataset, sono legati al genere e, in questo caso, c'è la possibilità di avere due canzoni che risultano simili per suddette caratteristiche, ma che appartengono a generi diversi.
Le variabili prese in esame riescono a distinguere bene gruppi musicali agli antipodi come Cantautorato e Heavy Metal, ma con generi come Rock e Punk non riescono a fare una distinzione netta, nonostante una persona possa apprezzare un genere e non essere interessata all'altro.

Fatte quindi queste considerazioni, una sfida per il futuro potrebbe essere quella di ricavare, mediante interrogazione dell'API Spotify, i generi musicali dei brani e stimare modelli di clustering e applicare tecniche di classificazione più complesse, che consentano di tenere in considerazione variabili qualitative, quali il genere.

FONTI

Fonte dataset: https://www.kaggle.com/datasets/bricevergnou/spotify-recommendation
Descrizione del Dataset: https://github.com/Brice-Vergnou/spotify_recommendation
Spotify Web API: https://developer.spotify.com/documentation/web-api