Report

Università di Pisa

Progetto di Data Mining 1

anno accademico 2022/2023

Ryerson Audio - Visual Database of Emotional Speech and Song (RAVDESS)

Students:

Giuseppe Palminteri

Graziano Amodio

Contents

1. Introduction

Il *Ryerson Audio - Visual Database of Emotional Speech and Song* (RAVDESS), è un dataset composto da registrazioni audio di 24 attori professionisti uniformemente divisi per genere. Gli attori registrano due diverse frasi con un accetto neutrale Nord-Americano. Agli attori, quindi, è stato chiesto di recitare e di cantare le frasi adottando diverse tipologie di espressività emozionali, registrando ogni frase sia ad una intensità normale che forte, aggiungendovi un'espressione neutrale. Lo scopo della seguente indagine è quello di analizzare il dataset per osservare eventuali correlazioni tra la lunghezza e l'intensità delle registrazioni per predire il tipo di interpretazione. Per fare ciò abbiamo suddiviso il lavoro in differenti fasi: Data Understanding, Data Preparation, l'implementazione di tre metodi di Clustering, Classificazione e Pattern ed Association Rules Mining.

2. Data Undestanding

2.1 Data Semantics

Il dataset si compone di 2452 records e di 38 features. Esso si riferisce a dati estrapolati direttamente dalle tracce audio del dataset (RAVDESS) che descrivono le caratteristiche principali delle registrazioni attraverso diverse variabili. Le variabili in questione si differenziano in categoriche e numeriche. Quelle categoriche indicano l'attore, il suo sesso, se si tratta di una traccia cantata o recitata, il testo tra le due diverse tipologie ed a quale ripetizione ci si riferisce. Sempre per questo macro-gruppo, per ogni record è indicato il tipo di emozione che l'attore deve interpretare, nello specifico per le frasi recitate all'attore è stato chiesto di interpretare le seguenti modalità espressive: calma, felicità, tristezza, rabbia, paura, sorpresa e disgusto. Lo stesso è stato fatto per le frasi cantate escludendo, però, le emozioni di sorpresa e disgusto. Infine, ogni espressione è stata registrata utilizzando sia un intensità forte che una intensità normale con l'aggiunta di una espressività della frase neutrale.

Di seguito una descrizione più dettagliata delle variabili categoriche:

1. **Modality**: si riferisce al tipo di traccia in questione. Il nostro dataset presenta solo modalità *"audio only"*.
2. **Vocal channel**: Indica se il testo viene cantato o recitato.
3. **Emotion**: Indica il tipo di emozione richiesta all'attore tra 7 possibili modalità per i testi recitati e 5 diverse possibili modalità per quanto riguarda i testi cantati.
4. **Emotional intensity**: Indica l'intensità nell'espressività richiesta all'attore tra normale, intensa (forte) oppure se si tratta di una intensità neutrale.
5. **Statement**: Indica la frase recitata tra due possibili: *"Dogs are sitting by the door"* oppurer *"Kids are talking by the door"*.
6. **Repetition**: Indica se la traccia audio in questione si riferisce alla prima o alla seconda ripetizione richiesta all'attore.

7. **Actor**: Indica a quale attore si riferisce la registrazione. Per ognuno è stato attribuito un numero per un totale di 24 attori.
8. **Sex**: Indica il sesso dell'attore. Ricordiamo che gli attori selezionati sono rappresentati in numero paritario tra i due generi.

Continuando la descrizione del nostro dataset, le seguenti variabili numerali descrivono delle caratteristiche tecniche del tipo di registrazione usata per le tracce audio. In particolare se si tratta di registrazioni mono o stereo, il numero di bit per campione audio, la frequenza della registrazione e il numero di bit per ogni frame.

1. **Channels**: Indica se le tracce audio sono monofoniche o stereofoniche.
2. **Sample Width**: Indica il numero di bit per ogni campione, questa è un informazione sempre relativa al tipo di registrazione utilizzata per le tracce audio. Ad 1 corrisponde 8 bit, mentre al valore 2 corrisponde 8 bit per campione.
3. **Frame rate**: Indica la frequenza in Hz usata per registrare i campioni audio.
4. **Frame Width**: Indica il numero di bytes per ogni frame.

Le successive variabili numeriche, sia discrete che continue, riferiscono informazioni in merito alla durata delle registrazioni e in merito ad alcune caratteristiche estrapolate dalle tracce audio.

Nel dettaglio parliamo di:

1. **Length**: indica la lunghezza delle tracce audio in millisecondi.
2. **Frame Count**: Indica la somma dei frames per campione audio.
3. **Zero Crossing Sum**: Indica la velocità con cui un segnale cambia da positivo-zero-negativo estrapolabile attraverso la loro somma.

Per ognuna delle seguenti variabili, per ogni record, sono stati estrapolati i valori della media*("mean")*, deviazione standard*("std")*, valore minimo*("min")* e massimo*("max")*, curtosi*("kur")* e indice di simmetria*("skew")*:

1. **Statistics of the original audio signal**: Caratteristiche statistiche principali delle tracce audio.
2. **Statistics of the Mel-Frequency Cepstral Coefficients**: Questa caratteristica descrive in modo conciso la forma complessiva di un inviluppo spettrale.
3. **Statistics of the Spectral Centroid**: Indica dove è situato il centro di massa in uno spettro. È quella caratteristica che aiuta a percepire ad esempio la brillantezza dei suoni.
4. **Statistics of the stft chromagram**: Indica le caratteristiche della trasformata di Fourier a tempo breve.

Di seguito riportiamo le 38 features distinte per attibuti:

- **variabili categoriche**: Emotion, Emotional Intensity, Actor, Sample Width, Frame rate, Frame Width,
- **variabili nominali binarie**: sex, channels, Repetition, Statement, modality, Vocal channel
- **variabili Numeriche discrete**: Length, Frame Count, Zero Crossing Sum,
- **variabili Numeriche continue**: Statistics of the original audio signal, Statistics of the Mel-Frequency Cepstral Coefficients, Statistics of the Spectral Centroid, Statistics of the stft chromagram.

2.2 Distribuzione statistica delle variabili

In questa fase analizziamo più approfonditamente la distribuzione statistica delle principali features.
Come è possibile osservare dal grafico (*fig:1*) le registrazioni con testi recitati si presentano di numero superiore in virtù del fatto che i records cantati non presentano le modalità *"surprised"* e *"disgust"*. Tali records sono distinti per sesso, caratteristica discriminante principale trattandosi di registrazioni vocali.

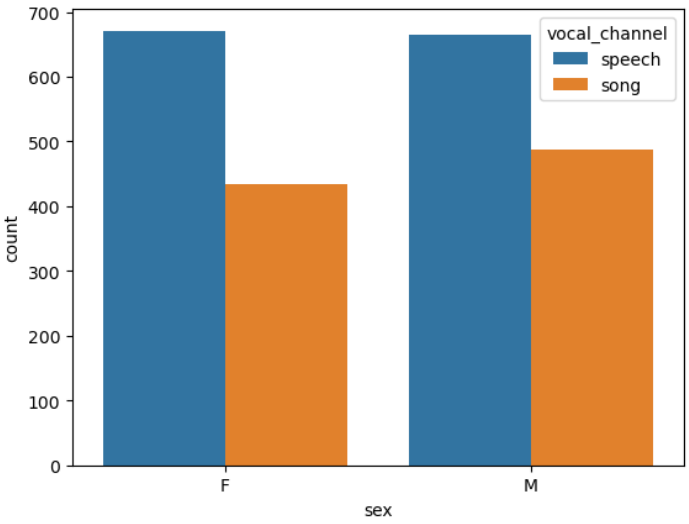

Figure 1: Composizione della classe "vocal channel"

È interessante notare come prendendo come riferimento la colonna *"lenght (ms)"* le canzoni sono visivamente più lunghe delle frasi recitate (*figure 2*) e che anche le emozioni presentano delle differenze che possono notarsi in questa prima fase di esplorazione. Come è possibile osservare a destra della (*figure 2*), le frasi recitate con una interpretazione *"surprised"* hanno una durata minore rispetto a tutte le altre e che, specularmente, quelle recitate con una interpretazione *"calm"* hanno una durata maggiore. Considerando che il numero di parole è lo stesso questo è un buon modo per individuare quali interpretazioni presentano la caratteristica di essere più veloci o lente.

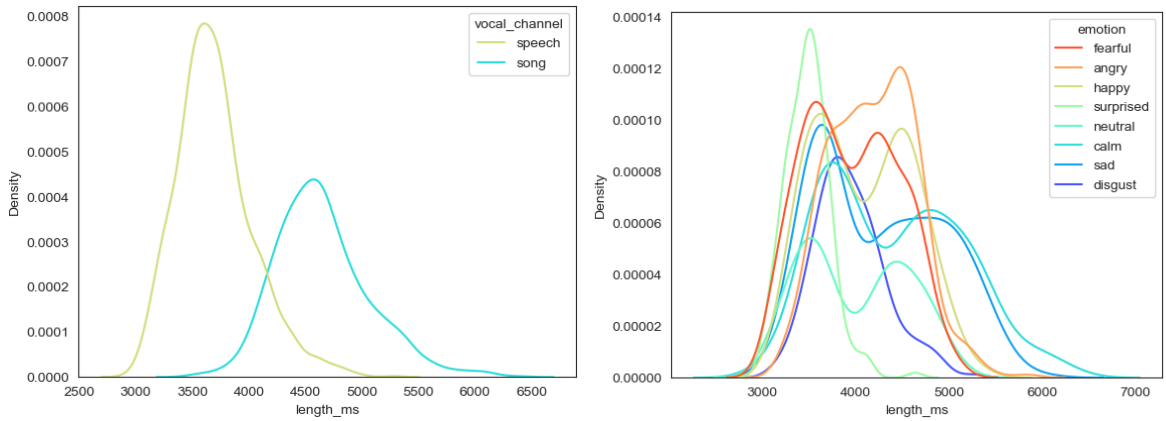

Figure 2: "distribuzione durata frasi recitate e cantate e distribuzione emozioni considerando la variabile length ms"

2.3 Preparazione del datset

In questa fase andiamo ad eliminare dal nostro dataset quelle features poco informative ai fini delle nostre analisi.

2.3.1 Colonne con incosistenze semantiche da eliminare

Esplorando il nostro dataset possiamo osservare che le distribuzioni statistiche di molte colonne non sono utili alle nostre analisi in quanto contengono lo stesso valore per ogni records, o quasi, e che per lo più informano sulle caratteristiche specifiche legate alle registrazioni
Nel dettaglio abbiamo pulito il dataset eliminando le seguenti colonne:

1. **modality**: Tutte le tracce audio presentano la modalità "audio only". Per questo motivo abbiamo eliminato questa colonna.
2. **frame rate**: Tutti i record hanno una frequenza di 48000 hz. È una caratteristica legata alla qualità della registrazione uguale per tutti i records.
3. **sample width**: Le tracce sono registrate ad una qualità di 16 bit. È una caratteristica legata alla qualità della registrazione uguale per tutti i record.
4. **stft max**: La colonna ha un valore unico per tutte le righe. Per questo abbiamo deciso di eliminare questa colonna.
5. **channels**: Tutte le registrazioni sono monofoniche. Quindi questa colonna detiene lo stesso valore per tutti i record.
6. **frame width**: I record sono registrati con lo stesso numero di bit per frame. Solo 6 tracce hanno un numero di frame doppio. Considerando le caratteristiche tecniche delle registrazioni pensiamo che in quel caso si tratti un errore e comunque essendo una caratteristica legata alla qualità di registrazione abbiamo pensato di eliminare questa colonna.

2.3.2 Missing values

Il dataset presenta alcune variabili con valori mancanti. In particolare la variabile *"actor"* presenta 1226 *missing values*. Osserviamo che possiamo ovviare al problema andando ad eliminare questa colonna in quanto presenta un numero di valori mancanti troppo grande(circa il 50%). Inoltre, ai fini della nostra analisi, considerando che il numero attribuito per ogni attore è riconducibile ad un ID dello stesso, possiamo farne a meno, soprattutto, se consideriamo di avanzare le successive fasi concentrandoci sulla distinzione per genere o sui tipi di interpretazione, ad esempio.
Parimenti, la variabile *"intensity"* presenta troppi *missing values*, 816 per l'esattezza, infatti, anche in questo caso abbiamo optato per eliminare l'intera colonna. Per giustificare questa scelta facciamo altresì affidamento ad una forte correlazione positiva (0,97) con la features *"mfcc min"*. Infine, un'altra colonna che presenta *missing values* è *"Vocal_channel"*. In questo caso abbiamo optato nel sostituire i 196 valori mancanti con la moda. Ritenevamo eccessivo eliminare l'intera colonna dinnanzi l'esiguo numero di "Nan" e per la sua importanza per le successive analisi.

2.3.3 Outliers

Successivamente abbiamo continuato la nostra fase di esplorazione andando ad individuare possibili valori anomali presenti nelle features del dataset. È possibile, infatti, che vi siano stati degli errori nell'estrapolare i dati dalle registrazioni audio.
Questo è evidente, ad esempio, nel caso di *"frame count"*, nel grafico seguente infatti è possibile osservare come siano presenti 35 records dal valore -1. Non potendo essere negativa la somma dei frame, i valori anomali sono stati sostituiti con il valore medio.

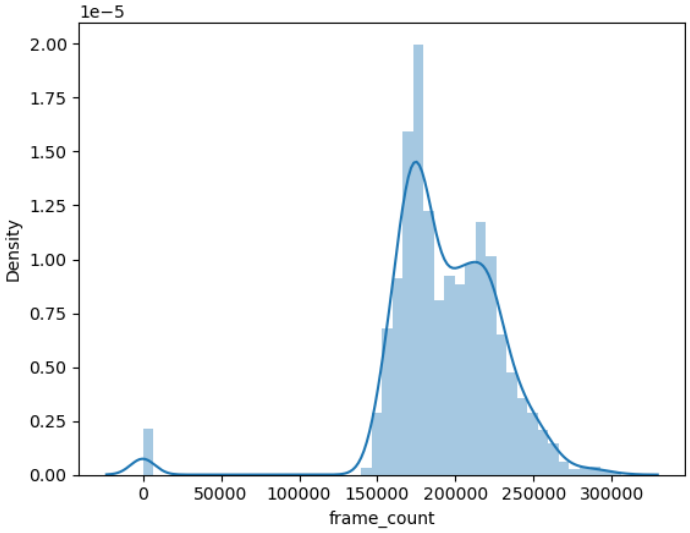

Figure 3: distribuzione "frame count"

Lo stesso processo di osservazione dei valori anomali e successiva sostituzione con il valore medio è stato effettuato anche per le variabili: ***stft-std, stft-skew, stft-kur***.

Continuando la nostra analisi, sempre grazie alle tecniche di visualizzazione, abbiamo optato per eliminare anche le variabili: **sc min e stft min**, infatti osservando le distribuzioni di queste features abbiamo notato come queste fossero completamente sbilanciate su un unico valore e come inoltre non presentassero correlazioni interessanti con le altre variabili.

2.4 Variabile trasformation

Per le variabili categoriche abbiamo assegnato ad ogni diversa modalità un numero intero mentre per quanto riguarda le variabili continue, ci siamo soffermati sulle seguenti : **zero-crossings-sum, frame-count e length-ms** applicando una trasformazione logaritmica, in quanto la loro distribuzione era fortemente caratterizzata da una asimmetria destra positiva. Questa applicazione ci ha permesso di visualizzare in maniera più agevole le distribuzioni delle suddette variabili.
Infine, attraverso la funzione *min-max-scaler* abbiamo deciso di normalizzare tutte le variabili continue affiché abbiano la stessa scala e, quindi, possano essere equiparate fra di loro. Di seguito rappresentiamo la distribuzione delle variabili dopo averle trasformate e normalizzate.

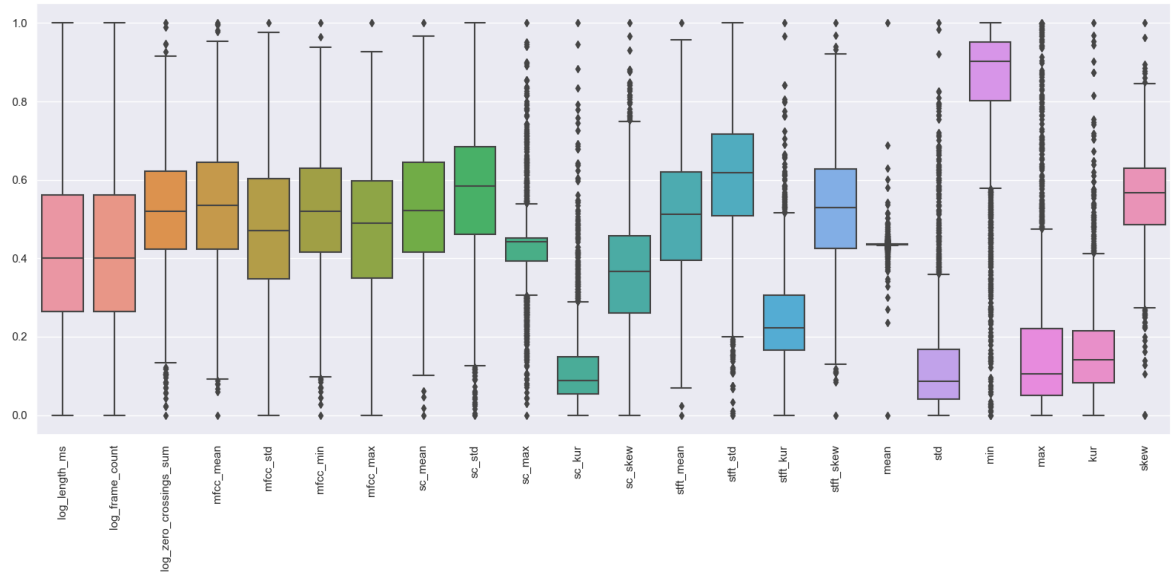

Figure 4: Boxplot

2.5 Analisi variabili maggiormente correlate

Per concludere la task del data understanding preparation abbiamo analizzato le correlazioni tra le variabili continue ed eliminato quelle superflue. La *figure 5* ci mostra le correlazioni esistenti tra le diverse variabili utilizzando l'indice di Spearman.

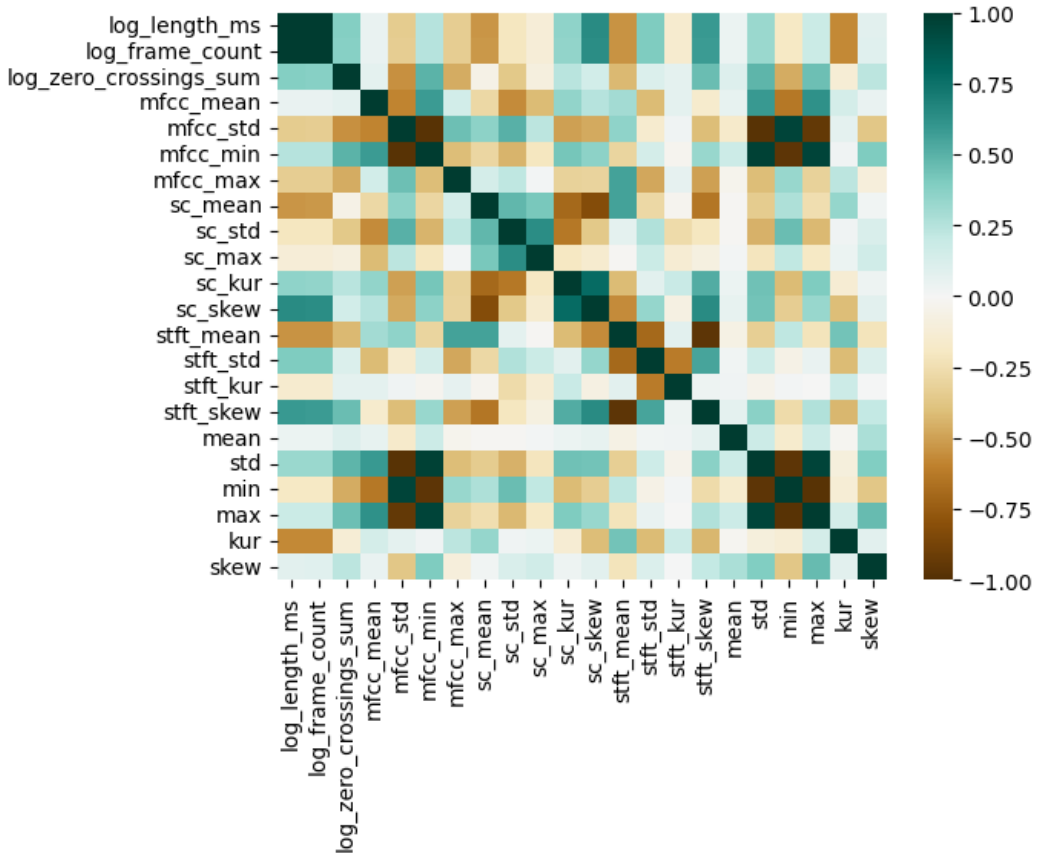

Figure 5: Correlation

Eliminazione delle variabili:

- **frame-count - length-ms**: Queste due variabili presentano una correlazione maggiore di 0,95 perché effettivamente rappresentano la stessa informazione. Infatti attraverso la somma dei frame otteniamo un informazione rispetto alla durata della registrazione.
- **max - std**: C'è una forte correlazione statistica probabilmente spiegabile con il fatto che le registrazioni con una più forte varianza hanno dei picchi più alti delle loro frequenze che ne caratterizzano una varianza, appunto maggiore considerando che il valore minimo più basso possibile è 0.
- **std - max - mfcc min**: La variabile mfcc_min presenta una forte correlazione con std e max (variabili a loro volta fortemente correlate).

Tenendo conto di quanto detto precedentemente, per diminuire la complessità del dataset, concludiamo la fase di data preparation eliminando la variabile **frame count** e le variabili statistiche **std** e **max**.

Adesso il nostro dataset è composto da sole 24 variabili, a differenza delle 38 iniziali, e costituisce un ottimo punto di partenza per le task successive.

3. Clustering

Abbiamo proseguito la nostra analisi attraverso lo studio dei cluster, ovvero i processi di raggruppamento di oggetti simili, all'interno del nostro dataset. Abbiamo deciso di iniziare questa task utilizzando il clustering per centroidi quindi utilizzando il K-means, successivamente il clustering gerarchico ed infine quello basato su densità cioè il DB-SCAN, quest'ultimo con una fase di pre-processing differente.

3.1 K-means

Il primo metodo utilizzato come già anticipato è il K-MEANS. I dati, già normalizzati con la funzione "min-max normalization", sono stati elaborati dall'algoritmo il quale richiede un solo parametro: K, ovvero il numero di clusters.

3.2 Scelta di k

Per scegliere il K ottimale per il nostro set di dati, abbiamo utilizzato il cosiddetto "Elbow method". Quindi, abbiamo cercato il miglior valore di k eseguendo il looping attraverso diversi valori di SSE e Silhouette score. Come si può osservare nella parte sinistra dalla figura 6 non c'è una forte curvatura nel grafico SSE per applicare il perfettamente il Elbow method, i risultati migliori si SSE si attestano tra 4 e 6 anche se in k=5 c'è un leggero segno di gomito. Il grafico della Silhouette score non ci aiuta molto in questa ricerca infatti non abbiamo notato un coefficiente rilevante tra i cluster.

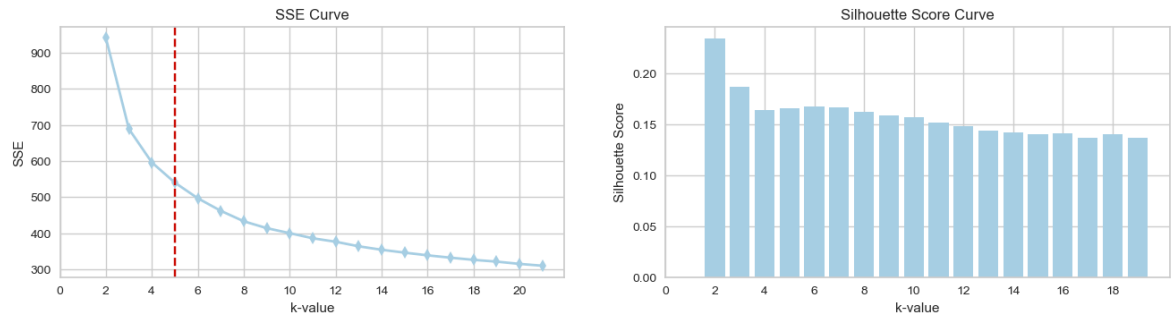

Figure 6: SSE Sihlouette score

In base ai risultati ottenuti si è deciso di scegliere K=5, con un SSE score 496.667 e una Sihlouette score di 0.165 che rappresenta il miglior compromesso tra il numero di clusters e il valore dei due coefficienti.

Di seguito la tabella in cui si indica la dimensione di ogni cluster:

Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
539	725	334	392	462

Grazie allo scatterplot, abbiamo una visualizzazione ideale della distribuzione spaziale dei cinque clusters prendendo in considerazione le features stft_skew e mfcc_mean.

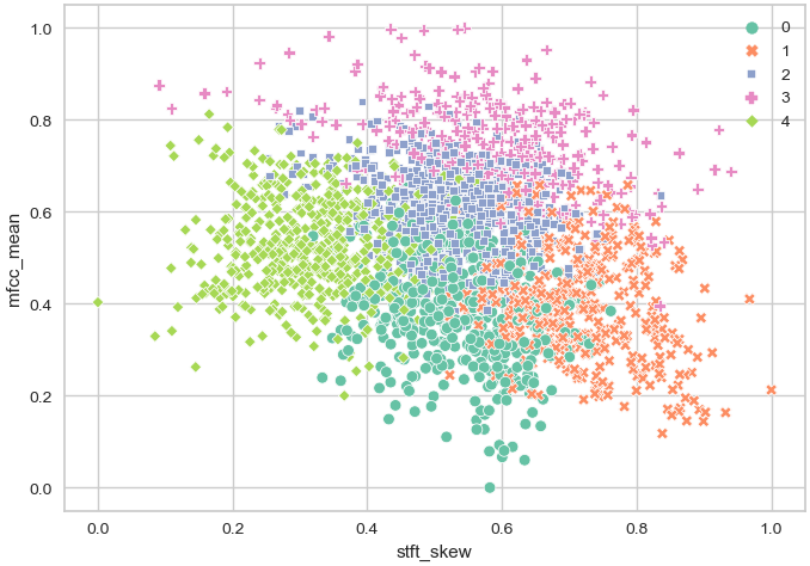

Figure 7: K-MEANS

3.3 Sintesi dei risultati

L'algoritmo appena descritto si rivela valido, ma non il migliore per questa task. I migliori risultati si sono ottenuti con le variabili "sex" e "vocal_channel". Nella parte sinistra della (*figure 8*) soffermandoci sulle coppie di clusters A-D e C-E, si può osservare rispettivamente la predominanza maschile e poi femminile del genere. Piu' nello specifico, se nei cluster A-D si osserva circa la totalità di recitazioni di genere maschile, nel cluster C ed E tale percentuale cambia abbassandosi notevolmente, evidenziando la netta prevalenza di "female".

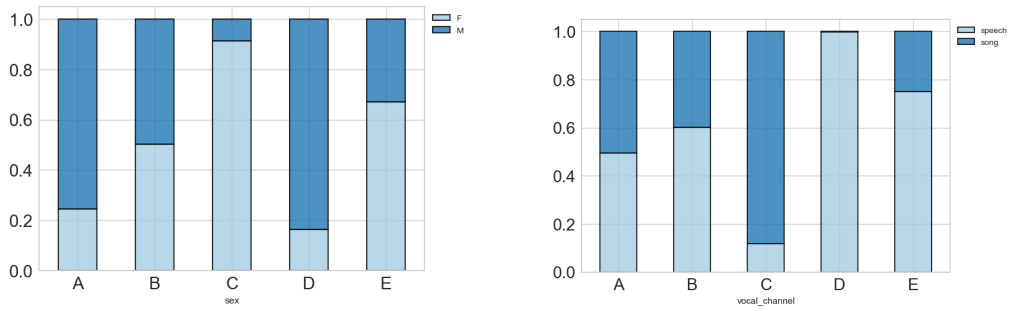

Figure 8: variabili sex e vocal channel

Nel grafico a destra, invece, si può osservare la distribuzione della variabile "vocal_channel", che ci indica se il testo viene cantato o recitato. Come ci aspettavamo, nella distribuzione della variabile all'interno dei cluster si può osservare la predominanza dell' attributo "speech". In particolar modo notiamo come il cluster D sia composto dalla totalità di tale attributo e il cluster E da circa il 75%.
Invece per quanto riguarda l'attributo song, la frequenza maggiore la ritroviamo principalmente nel cluster C, dove registriamo circa il 90%, mentre per i primi due clusters si rimarca una situazione intermedia.

3.4 Hierarchical

Il secondo metodo di clustering utilizzato è quello gerarchico. Le verifiche sono state condotte applicando la tipologia Agglomerative sulla base della definizione di distanza Euclidea e di clustering proximity (Single, Average, Complete e Ward).
Inizialmente è stato utilizzato il Single linkage il quale si è dimostrato il peggiore per il nostro studio in quanto raggruppa piu' di 2200 punti in un unico cluster. Si osserva infatti quanto le distanze, tra ognuno di essi siano molto simili, tanto che la radice dell'albero dista dal precedente raggruppamento soltanto di 0.1, il che vuol dire che il distacco dei sottostanti non sarà mai maggiore di tale valore.
Per questa ragione non si è ritenuto opportuno svolgere ulteriori analisi. Il motivo di ciò è la rilevante densità dei punti del dataset che essendo così vicini non permettono una buona clusterizzazione con il Single method.

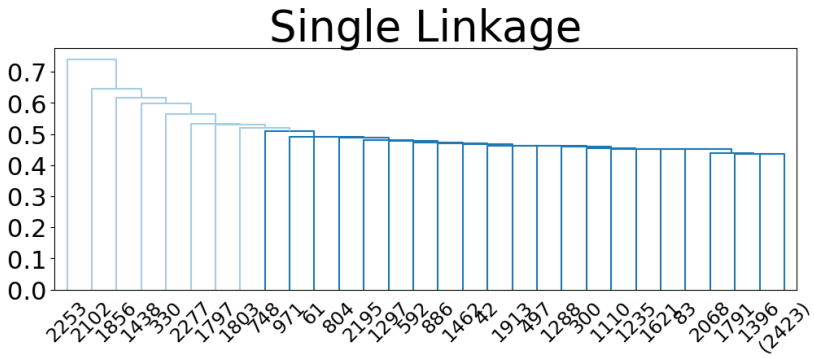

Figure 9: vocal-channel

Si è quindi proceduto utilizzando l'Average method Linkage, che risulta migliore, ma comunque non fornisce una differenziazione dei cluster chiara e distinta. Una visualizzazione simile, si è ottenuta con il Complete method Linkage, da cui si possono trarre conclusioni analoghe a quelle precedenti. Infine, per progressiva miglioria di clusterizzazione, utilizzando l'approccio gerarchico, i risultati piu' efficaci sono stati rilevati tramite il Ward's method Linkage, che riportiamo di seguito, il quale si basa sull'agglomerazione dei gruppi in termini incrementali di SSE (analogo al K-Means):

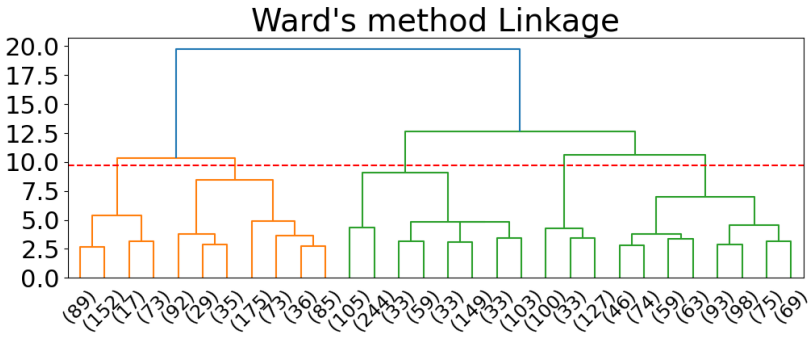

Figure 10: Hierarchical

Come si può osservare nella (*figure 10*), come per il K-MEANS, il dataset viene suddiviso in 5 cluster ben distinti. Tale scelta è stata fatta perchè i valori piu' alti di Silhouette score si attestano per k=5, con un punteggio di circa 0.166 (*figure 11*); confermando anche la scelta di K del K-MEANS oltre a quella ottenuta tramite il suddetto metodo.

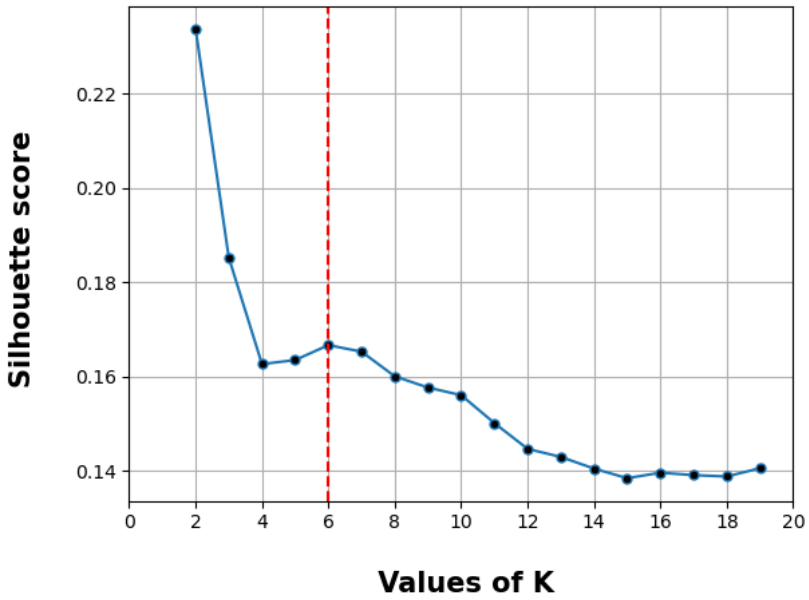

Figure 11: Hierarchical

3.5 Sintesi dei risultati

Anche in questo caso abbiamo analizzato la dimensione dei cinque cluster e successivamente le variabili che meglio li discriminavano:

Cluster A	Cluster B	Cluster C	Cluster D	Cluster E
911	449	328	294	470

Come nel K-MEANS si è evinto che tra le variabili inserite in input, anche questa volta, quelle che ci hanno restituito risultati migliori sono sex e vocal_channel. Per

quanto riguarda quest'ultima, come possiamo vedere nel grafico a sinistra della (*figure 12*), la coppia di cluster B-E è composta quasi esclusivamente dal labels "speech" mentre il cluster D da "song" per circa il 90%. Risultati leggermente peggiori si riscontrano per la variabile sex, grafico a destra, dove i cluster B e D sono composti principalmente dal genere femminile mentre il cluster E da quello maschile.
Questo algoritmo si confermerà il migliore tra quelli sviluppati per questa task. Possiamo concludere dicendo che le variabili che piu' discriminano il nostro dataset sono sex e vocal_channel. Infatti, quest'ultima variabile sarà la nostro riferimento per la task successiva, ovvero la classification.

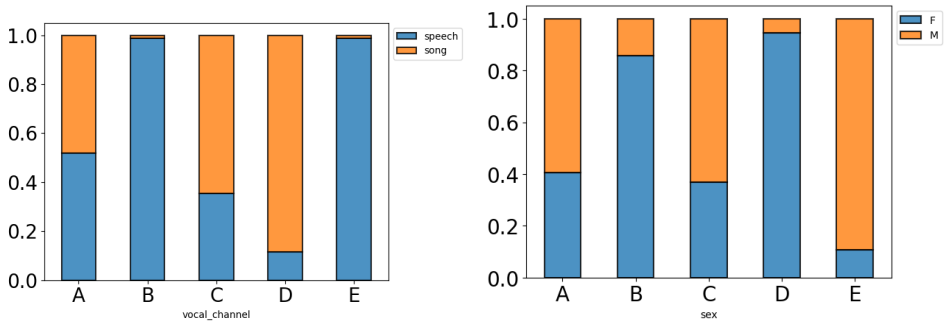

Figure 12: Hierarchical

3.6 DBSCAN

Il DBSCAN è il terzo ed ultimo algoritmo che è stato implementato ai fini della nostra ricerca; si tratta di un metodo density-based che permette di individuare il numero dei cluster di un dataset. Si ricorda che esso lavora in modo da determinare se il numero minimo di punti siano abbastanza vicini l'uno dall'altro da essere considerati parte di un singolo cluster. Per questa analisi abbiamo attuato una fase di pre-processing differente rispetto agli altri algoritmi di clustering; tenendo conto che questo algoritmo ha il vantaggio di riconoscere bene gli outliers, abbiamo utilizzato un dataset piu' 'sporco', cioè non abbiamo attuato quel processo di sostituzione degli outliers con il valore medio delle variabili, come indicato precedentemente nella fase di data understanding.

3.7 Scelta dei parametri

In primo luogo, tramite la visualizzazione grafica del KNN algorithm (*figure 13*) si evince che il valore ottimale di EPSILON corrisponde a 0.38 e questo vale per k compreso tra 2 e 5 inclusi. In seguito, massimizzando il Sihlouette Score si è trovato il numero ottimale di min_sample che risulta essere 2, con un punteggio di 0.232.

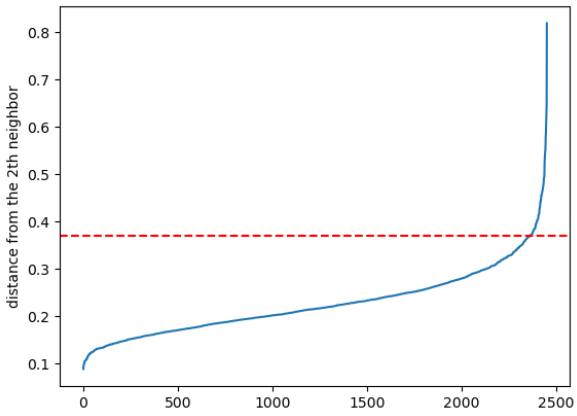

Figure 13: DBSCAN

Approfondendo l'analisi del DBSCAN, si evince che la composizione dei due clusters è nettamente sbilanciata, infatti, un cluster è composto quasi dalla totalità dei records per l'esattezza 2399, contro i soli 53 dell'altro.

3.8 Sintesi dei risultati

Purtroppo, nonostante siano stati presi in considerazione i parametri migliori possibili, non siamo riusciti ad ottenere risultati rilevanti. Infatti, possiamo concludere la task del clustering affermando che quest'ultimo algoritmo non risulta efficiente per il nostro dataset. Risulta logico sostenere che, dal momento in cui è presente una sola componente che comprende il 98% dei dati, non si riesca a discriminare bene il dataset; tutti i dati, in pratica, si "clusterizzano" in uno solo, questo è dovuto alla tipologia dei dati e alla loro densità. Il modello pertanto non riesce a dividere correttamente i punti perchè si trovano molto vicini e quindi l'algoritmo non riesce a discriminarli.

4. Classification

La classificazione è un metodo di machine learning supervisionato che ci permette di assegnare a delle classi definite, oggetti presenti all'interno del nostro dataset. Questa metodologia è usata per classificare variabili categoriche e nel nostro caso la useremo per classificare la variabile "vocal channel".

4.1 Preparazione del dataset

Per il funzionamento dei modelli di classificazione utilizziamo il dataset precedentemente "pulito" nella fase di data understanding, in quanto non presenta *missing value* o altri problemi che necessitano una ulteriore fase di *pre-processing*.
Successivamente si è proceduto a trasformare le variabili categoriche in variabili *dummy*, quindi che si esprimono con 0 e 1.
Il dataset è stato diviso in *training set* 70% e in *test set* 30%.
La variabile target (*vocal channel*) presentava uno sbilanciameno tra le modalità *speech* e *song*, data dal fatto che per la modalità *song* non vi sono registrazioni per le

emozioni *surprised, disgust*. Per ottenere delle prestazioni più efficaci dai classificatori, constatando una situazione di *overfitting*, abbiamo bilanciato il dataset attraverso lo SMOTE (Synthetic Minority Oversampling).
Per la regolazione dei parametri abbiamo utilizzato il metodo Grid Search e la Cross Validation, per la quale abbiamo considerato l'addestramento composto dal set di validazione e dal set di addestramento.

4.2 Decision Tree

Il primo modello di classificazione utilizzato è il Decision Tree, modello che si verificherà il più performante per la nostra analisi.
Attraverso il grid_searh_estimator abbiamo sintonizzato il nostro classificatore con parametri più efficaci di quelli standard, ottenendo miglioramenti nei risultati.
In particolare:

Paramentri	Valori
criterion	entropy
min sample leaf	0.001
min sample split	0.01
max depth	6

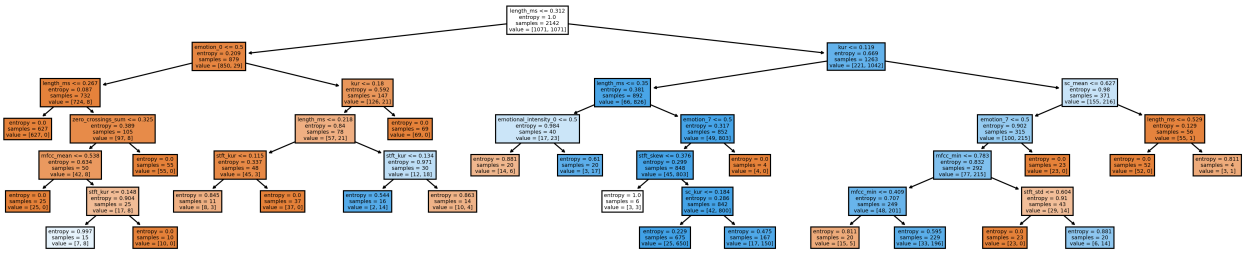

Figure 14: Decision Tree con parametri estratti con Grid Search e CV

```
=== CLASSIFICATION REPORT ===
              precision    recall  f1-score   support

      speech       0.96      0.88      0.92       460
        song       0.83      0.94      0.88       276

    accuracy                           0.90       736
   macro avg       0.89      0.91      0.90       736
weighted avg       0.91      0.90      0.90       736
```

Figure 15: Classification report Decision Tree

Dalla tabella del *classification report* del Decision Tree si evince come il classificatore sia preciso nel classificare le due diverse modalità di *vocal channel*. In particolare l'*f-1 score*, per entrambe le modalità è molto elevato.

4.3 Interpretazione Decision Tree

Il miglior risultato per l'applicazione del Decision Tree risulta quello che utilizza il dataset ottenuto attraverso la SMOTE applicando come *gain criterio* l'entropia.
Il primo *split* che divide esattamente a metà il dataset si basa sulla variabile *length ms*, che risulta essere la variabile più diffusa lungo l'intero albero decisionale. Gli split successivi si basano maggiormente sull'utilizzo delle variabili *emotion* e *kur*.

4.4 Naive Bayes

Il secondo algoritmo utilizzato per questa task è il Naive Bayes. Questo algoritmo di classificazione si rivelerà il peggiore tra quelli utilizzati in termini di prestazioni. Infatti si rivela il meno adatto per il nostro tipo di dataset.

```
=== CLASSIFICATION REPORT ===
              precision    recall  f1-score   support

      speech       0.99      0.45      0.61       460
        song       0.52      0.99      0.68       276

    accuracy                           0.65       736
   macro avg       0.75      0.72      0.65       736
weighted avg       0.81      0.65      0.64       736
```

Figure 16: Classification report Naïve Bayes

Osservando il *classification report* del Naïve Bayes si evince, osservando i diversi valori di *precision* e *recall* che sovrastima molto i falsi positivi, quindi nel nostro caso le *song* individuate come *speech*.

4.5 KNN

L'ultimo modello di classificazione utilizzato è il K-Nearest Neightbor (KNN). Anche in questo caso è stata usata la Grid Search per sintonizzare i parametri più efficaci e nel dettaglio, per migliorare le prestazioni del classificatore è stato effettuato un *over-sampling* del dataset.
I parametri scelti dalla Grid Search sono:

Paramentri	Valori
algorithm	ball tree
n neighbors	9
distance	Manhattan
weights	distance

```
=== CLASSIFICATION REPORT ===
              precision    recall  f1-score   support

      speech       0.96      0.88      0.92       460
        song       0.83      0.94      0.88       276

    accuracy                           0.90       736
   macro avg       0.89      0.91      0.90       736
weighted avg       0.91      0.90      0.90       736
```

Figure 17: Classification report KNN

Osservando il *classification report* del KNN si nota come questo classificatore con il nostro dataset ottiene risultati molto vicini al Decision Tree.

4.6 Sintesi dei risultati e miglior modello di previsione

Una valutazione di comparazione tra i diversi modelli di classificazione usati è stata effettuata attraverso la ROC curve per ognuno. Considerando le performance riscontrate il Decision Tree si rivela migliore ma le prestazioni del KNN sono molto simili. Il Naive bayes invece, risulta il peggiore soprattutto perché va a considerare un numero di False Positive troppo elevato.
Sempre a proposito del *Naïve Bayes*, osservando la ROC curve ed in particolar modo la *microaverage curve*, notiamo come appunto il numero di falsi positivi è maggiore rispetto gli altri classificatori, che attestano valori di *micro e macro average* allineati con le curve delle classi.

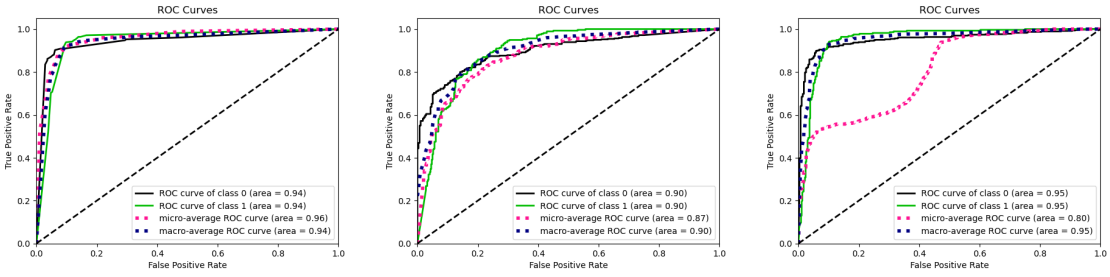

Figure 18: ROC Curve in ordine: DT, KNN e NB

5. Pattern Mining

In questa sezione finale sono individuate ed analizzate le regole di associazione trovate nel dataset.*lenght ms* è l'attributo posto sotto maggiore attenzione in questa fase.

5.1 Preparazione del dataset

Per l'analisi sono state prese in considerazione le seguenti variabili:

- **vocal channel**: variabile categorica
- **emotion**: variabile categorica
- **length ms**: Variabile continua discretizzata in 3 parti: *"short, medium e long length"*.
- **mfcc min**: variabile continua discretizzata in 3 parti *"low, medium e high intensity"*. Abbiamo, considerando la correlazione positiva con la variabile *intensity* maggiore del 0.9, usato *mfcc min* per discretizzare con tre livelli di intensità.

Come è possibile notare, ci siamo voluti soffermare nel cercare dei pattern che abbiano a che fare con le maggiori caratteristiche udibili dagli umani per trovare delle regole di associazione. Anche la scelta di discretizzare le variabili su tre livelli rispecchia questa decisione.
Questa scelta ci ha portati ad escludere molte variabili successivamente a molte prove effettuate ottenendo un numero maggiore di *frequent pattern* ma privi di significatività per un analisi sulle regole di associazione.

5.2 Preparazione del dataset Estrazione dei frequent itemset e valutazione del supporto minimo

Iniziando la nostra analisi abbiamo ricercato il parametro migliore di supporto minimo facendolo variare su un intervallo compreso tra 1 a 20 confrontandolo con il numero di *frequent pattern* che si andavano ad osservare.

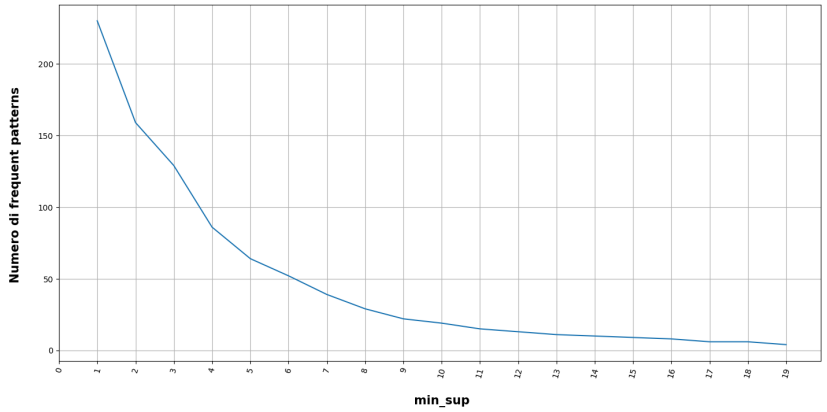

Figure 19: Frequent patterns per numero di supporto minimo scelto

Il numero di supporto minimo scelto è stato 15 alla quale corrispondono 9 frequent pattern.
Il ragionamento alla base di questa scelta è stato dettato dal fatto che scendere al di sotto di questa soglia rischiava di restituirci un numero elevato di pattern senza nessun riscontro significativo. I risultati sono stati ritenuti positivi in questo senso perché si nota che con un supporto del 33%, agli *speech* corrisponde una *short length* e con un supporto del 27% alle *song* corrisponde una *long lenght*. Quindi nel nostro dataset, soffermandoci sulla variabile *vocal channel*, ampiamente approfondita in questo lavoro, vi sono delle conseguenze interessanti in termini di *pattern* con la lunghezza delle note audio in considerazione.
Queste considerazioni sono frutto di numerosi tentativi con diversi numeri di supporto minimo e diversi valori di *item* minimi considerati. In virtù di questo, ad un numero di support minore del 10% ottenevamo un numero di *frequent items* pari a 19 che, nel nostro caso, considerando le poche variabili prese in considerazione abbiamo ritenuto non utili. L'ideale è considerare come soglia di supporto minimo il 25% , per ottenere significatività dato il nostro dataset preso in considerazione. Per quanto riguarda il numero di *item* minimo, con valori maggiori di 2 non abbiamo ottenuto nessun risultato utile.

5.3 Estrazione delle Association Rules con diversi valori di confidence e discussione delle regole piu‘ interessanti

Il primo passo per una analisi volta alla ricerca di regole di associazione è sicuramente osservare, al variare del livello soglia di *confidence*, il numero di associazioni come ci mostra il grafico *fig: 25*.

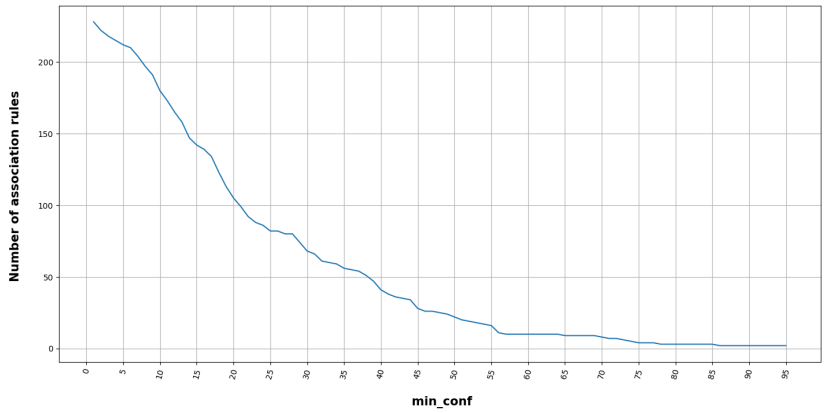

Figure 20: numero di association rules per diversi valori di confidence

Il nostro obiettivo però è di considerare un livello di *confidence* superiore almeno all'80%. Nel grafico in *fig: 26*, si osserva chiaramente che fino ad un livello di *confidence* dell' 86% sono presenti 3 regole di associazione, oltre questa soglia ne restano soltanto 2 fino ad un livello di confidenza del 100%
Questo risultato lo abbiamo reputato comunque positivo in relazione alle poche variabili prese in considerazione perché ci mostra una chiara regola di associazione nel dettaglio che lega il fatto che le canzoni abbiano registrazioni più lunghe delle frasi recitate.

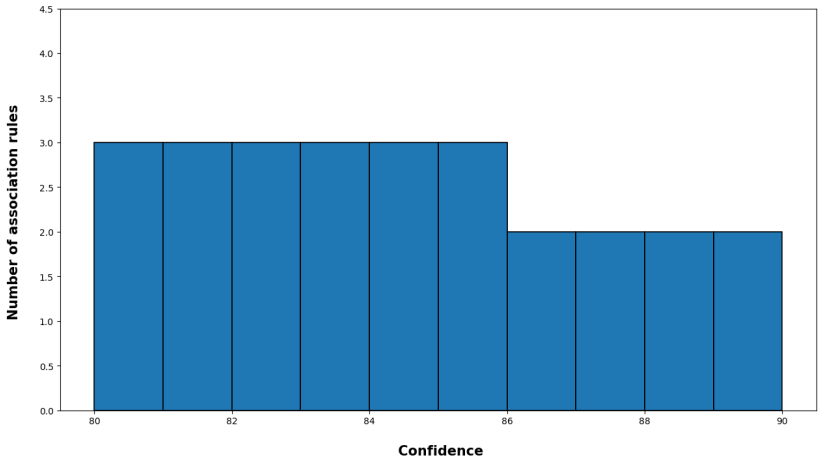

Figure 21: Dettaglio per livelli di confidence compresi tra 80 e 90

Osservando anche altre configurazioni, otteremmo un numero elevato di regole di associazione solamente considerando una *confidence* intono al 50% e un numero minimo di supporto troppo basso per ottenere dei dati che abbiano una significatività in termini di analisi.
Attraverso un istogramma sono stati analizzati i diversi livelli di lift al variare del numero di Association rules. I parametri scelti rispecchiano quanto detto precedentemente, perciò il suppporto è pari al 15% e il livello di confidenza è pari a 80.
Con questi parametri sapevamo già che ci fossero 3 diverse regole di associazione. La cosa interessante è notare come tra queste, una AR ha un lift value molto elevato, mostrandosi come più interessante rispetto alle altre. La regola in questione afferma che se le registrazioni sono lunghe allora si tratta di canzoni. Le altre, complementari a questa, affermano che se la registrazione è corta allora si tratta di registrazioni vocali registrate.

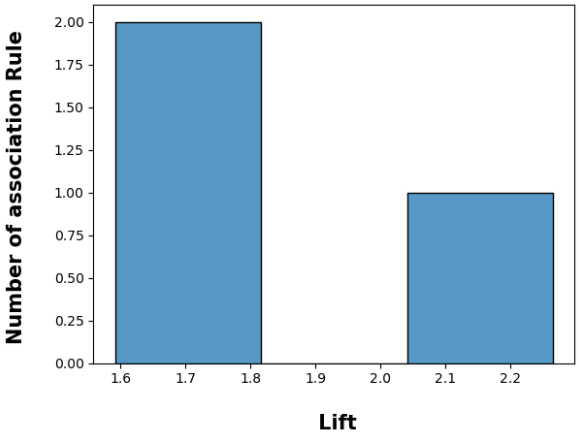

Figure 22: numero di association rules per i valori di lift corrispondenti