
Progetto Machine Learning

SPOTIFY AWARD

GIANLUCA GIUDICE - 830694

MARCO GRASSI - 829664

Contents

1	Introduzione	3
1.1	Descrizione del problema	3
1.1.1	Spotify	3
1.1.2	Premi per i singoli musicali	3
1.2	Approccio al problema	4
1.3	Struttura del codice	4
2	Dataset	5
2.1	Acquisizione del dataset	5
2.1.1	Dataset spotify - Kaggle	5
2.1.2	Premi canzoni - Wikipedia	5
2.2	Descrizione del dataset	7
2.2.1	Spotify	7
2.2.2	Premi	8
2.3	Data integration	8
2.3.1	Record linkage con MongoDB	8
2.4	Analisi esplorativa	9
2.4.1	Assunzioni sul dominio	9
2.4.2	Distribuzione dei valori	12
2.4.3	Artisti nelle canzoni	15
2.4.4	Correlazione features	15
2.4.5	Principal component analysis	15
2.4.6	Creazione dataset bilanciato	15
2.5	Scelta delle features	15
2.5.1	Variabili scartate	15
2.5.2	Rappresentazione BOW degli artisti	15
2.5.3	Variabili categoriche	15
2.5.4	Coordinate PCA	15
2.5.5	Normalizzazione e standardizzazione	15
2.5.6	Dataset proiettato componenti principali	15
2.5.7	Riassunto feature finali utilizzate	15
3	Campagna sperimentale	16
3.1	Approccio	16
3.1.1	10-folds cross validation	16
3.1.2	Training set e test set	16
3.1.3	Model selection	16
3.2	Misure di performance	16
3.2.1	Accuracy	16
3.2.2	Precision, Recall e F-measure	16
3.2.3	Curve ROC	16

<i>CONTENTS</i>	2
3.3 Support Vector Machine	16
3.3.1 Kernel	16
3.4 Decision Tree	16
3.5 Esperimenti	16
3.5.1 Performance	16
3.5.2 Modelli a confronto	16
4 Conclusioni	17

Chapter 1

Introduzione

In questo capitolo viene introdotto il problema, il dominio di riferimento e l'approccio adottato per la risoluzione.

1.1 Descrizione del problema

In questo elaborato consideriamo i singoli musicali. Al giorno d'oggi le canzoni vengono spesso ascoltate dagli utenti tramite piattaforme di streaming musicale.

L'obiettivo del lavoro è quello di analizzare le features dei singoli musicali così da prevedere se una canzone diventerà o meno di successo. Nella sezione successiva verrà meglio specificato cosa si intende per **singolo musicale di successo** (subsection 1.1.2).

1.1.1 Spotify

Spotify è un servizio di riproduzione digitale in streaming di musica, podcast e video, con accesso immediato a milioni di brani e altri contenuti di artisti provenienti da tutto il mondo. Questa piattaforma viene utilizzata da milioni di utente per ascoltare canzoni e nello specifico singoli musicali.

Da questo servizio è possibile ottenere migliaia di brani musicali, infatti Spotify mette a disposizione una API da cui è possibile scaricare informazioni su questi brani con associate alcune caratteristiche. Pertanto oltre alle classiche informazioni di un brano come "titolo" o "artisti" si avrà a disposizione una serie di caratteristiche specifiche del brano, ad esempio quanto una canzone è "energica" o "ballabile".

1.1.2 Premi per i singoli musicali

Come vedremo in subsection 2.2.1 il dataset ottenuto da Spotify mette a disposizione un campo "popularity" che indica quanto può essere considerata popolare. Tuttavia questa caratteristica non ci sembra adeguata per identificare una canzone come di successo oppure no.

Per questo motivo consideriamo una canzone di successo in base alle certificazioni ottenute, rispettivamente "disco d'oro" o "disco di platino". Queste premi sono dei riconoscimenti vinti da un brano musicale e storicamente fanno riferimento al numero di copie vendute da un singolo. Tuttavia con la costante crescita dell'utilizzo di servizi per lo streaming di brani musicali, da qualche anno questi riconoscimenti vengono assegnati anche considerando il numero di streaming sulle diverse piattaforme, tra cui Spotify.

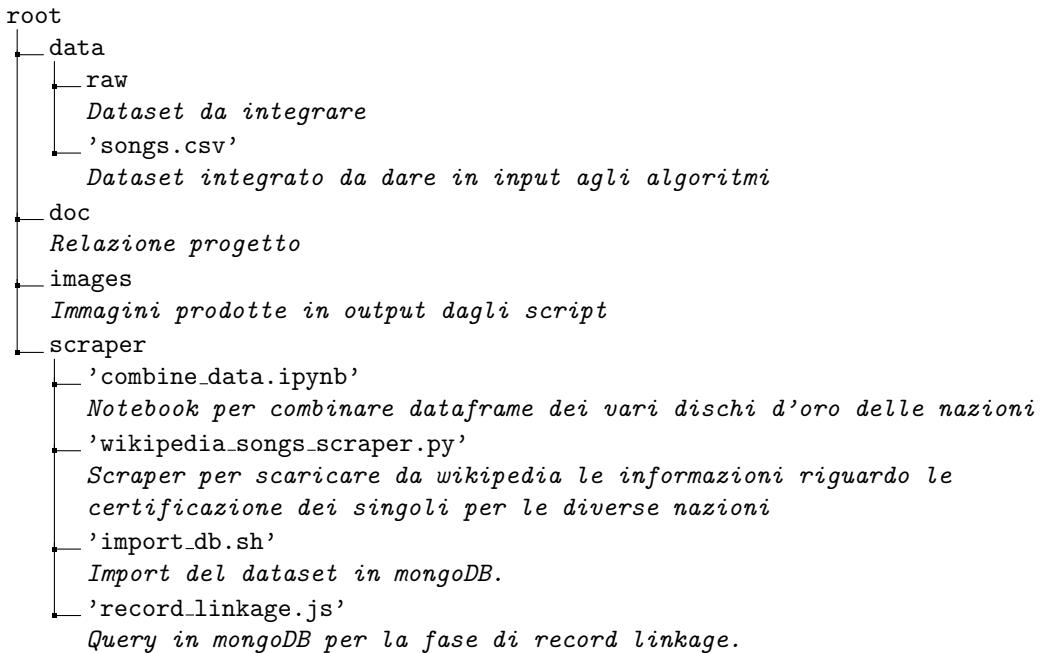
Riteniamo che questo riconoscimento sia una metrica oggettiva per considerare un singolo come di successo.

1.2 Approccio al problema

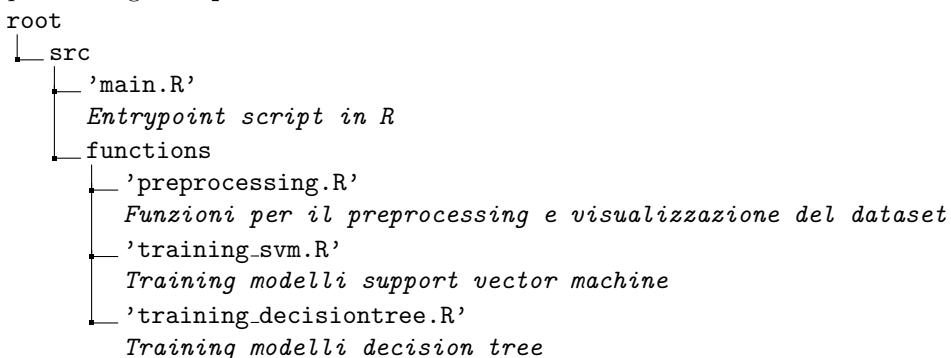
Il problema viene approcciato come un **task di classificazione binaria**. Dato un singolo musicale vogliamo prevedere se questo sarà di successo o no. Pertanto sviluppiamo modelli supervisionati di machine learning partendo da un dataset etichettato, in questo modo è possibile classificare i brani musicali.

1.3 Struttura del codice

Di seguito viene brevemente spiegata la struttura del codice, inoltre viene sotto indicata la working directory e l'entry point del programma.



In particolare gli **script in R** si trovano in:



N.B.: Per come è stato progettato il codice, la working directory è la `root/` e non la cartella `src/`. L'**entry point del programma** è lo script `src/main.R`

Chapter 2

Dataset

In questo capitolo viene analizzato il dataset. Viene quindi spiegato da dove è stato preso il dataset, descritte le covariate e viene effettuata un'analisi esplorativa.

2.1 Acquisizione del dataset

Il dataset completo contenente sia le informazioni sui singoli musicali che riguardo le varie certificazioni, ovvero le etichette del dataset, non è disponibile da una sola sorgente. Pertanto abbiamo ottenuto le informazioni necessarie da diversi sorgenti per poi integrare i dati.

2.1.1 Dataset spotify - Kaggle

Per quanto riguarda i brani con le relative informazioni e caratteristiche del brano, Spotify mette a disposizione un'API da cui si possono ottenere questi dati.

Con questa tecnica sono stati ottenuti i dati relativi ai brani, un utente ha quindi caricato il dataset sul sito kaggle.com. Il dataset è disponibile su kaggle all'url indicato¹.

Il dataset contenente queste informazioni è il file: `data/raw/to_integrate/data.csv`

2.1.2 Premi canzoni - Wikipedia

Le informazioni riguardo i premi delle canzoni, ovvero le varie certificazioni vinte, sono disponibili su wikipedia.org. Questa informazione costituisce di fatto l'etichetta per classificare ogni brano musicale.

Nello specifico siamo interessati a:

- Singoli certificati oro
- Singoli certificati platino
 - 1. Singoli certificati 1 volta platino
 - 2. Singoli certificati 2 volte platino
 - 3. ...
 - 4. Singoli certificati N volte platino

Inoltre le varie certificazioni dei singoli vengono considerati in base al paese. I paesi da noi presi in considerazione sono:

¹**Dataset Spotify:** <https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks>.

- Italia
- Australia
- Stati Uniti d'America
- Regno Unito
- Canda
- Danimarca

Si noti come una certificazione può essere consegnata in diversi paesi alla stessa canzone.

Certificazioni disco d'oro

Per quanto riguarda i dischi d'oro, facciamo riferimento a questa pagina su wikipedia². Fissato quindi uno stato (ad esempio l'Italia) è possibile visualizzare la pagina contenente la lista dei singoli che hanno vinto quel particolare premio. La lista è un elenco di url che puntano alla pagina wikipedia della canzone, un esempio a questo url³.

Certificazioni disco di platino

Ragionamento analogo viene fatto per i dischi di platino, con l'unica differenza che la pagina è questa⁴, inoltre dal momento che i singoli posso vincere più volte un disco di platino, si considerano non solo i singoli che hanno vinto una volta il disch di platino ma anche quelli che l'hanno vinto N volte.

Scraping

Ottenuti quindi i puntatori alle canzoni certificate, è possibile accedere alla pagina wikipedia del singolo, la quale contiene una tabella riassuntiva della canzone. Un esempio di tabella viene mostrato nella Figura 2.1.

²**Disco d'oro per stato:**

https://it.wikipedia.org/wiki/Categoria:Singoli_certificati_oro_per_stato.

³**Singoli certificati disco d'oro in Italia:**

https://it.wikipedia.org/wiki/Categoria:Singoli_certificati_disco_d%27oro_in_Italia.

⁴**Singoli certificati disco di platino per stato:**

https://it.wikipedia.org/wiki/Categoria:Singoli_certificati_platino_per_stato.

Chico (singolo)

Da Wikipedia, l'enciclopedia libera.

Chico è un singolo del rapper italiano Gué Pequeno, pubblicato il 31 luglio 2020 come secondo estratto dal sesto album in studio *Mr. Fini*.^[2]

Indice [nascondi]	
1	Classifiche
1.1	Classifiche settimanali
1.2	Classifiche di fine anno
2	Note
3	Collegamenti esterni

Classifiche [[modifica](#) | [modifica wikitesto](#)]

Classifiche settimanali		Classifiche di fine anno	
[modifica modifica wikitesto]		[modifica modifica wikitesto]	
Classifica (2020)	Posizione massima	Classifica (2020)	Posizione
Italia ^[3]	5	Italia ^[4]	10

Chico	
Artista	Gué Pequeno
Featuring	Rose Villain e Luchè
Tipo album	Singolo
Pubblicazione	31 luglio 2020
Durata	3:33
Album di	<i>Mr. Fini</i>
provenienza	
Genere	Pop rap
Etichetta	Island
Produttore	Sixpm
Formati	Streaming
Certificazioni	
Dischi di platino	ITALIA (3) ^[1] (vendite: 210 000+)
Gué Pequeno - cronologia	
Singolo precedente	Singolo successivo
<i>Saigon</i> (2020)	<i>Bla Bla</i> (2020)
Luchè - cronologia	
Singolo precedente	Singolo successivo
<i>Come me</i> (2019)	<i>Maserati (Reloaded)</i> (2020)

Figure 2.1: Esempio di tabella per un singolo musicale su wikipedia

Viene quindi creato uno script per effettuare scraping delle tabelle, il codice è nella directory `scraping/wikipedia_songs_scraper.py`.

Successivamente si integrano le informazioni dei diversi stati, e tipi di certificazione vinti. Il risultato di questa operazione è il file `data/raw/to_integrate/awards_cleaned.csv`.

2.2 Descrizione del dataset

In questa sezione vengono elencate e descritte le features del dataset. Trattandosi di un problema supervisionato, ad ogni istanza viene associata la corretta etichetta, ovvero la classe positiva per i brani musicali che hanno vinto un premio e classe negativa per i brani che non hanno vinto un premio.

2.2.1 Spotify

Di seguito viene descritto il dataset proveniente da kaggle, ovvero quello contenente le informazioni dei brani.

ATTRIBUTO	DESCRIZIONE	TIPO
id	Identificativo della canzone (generato da spotify)	Intero
name	Titolo della canzone	Stringa
artists	Lista degli artisti che compaiono nel brano	Stringa
year	Anno del brano	Intero
duration_ms	Durata della traccia in millisecondi	Intero
acousticness	Metrica riguardante quanto un brano risulta "acustico"	Float [0, 1]
danceability	Metrica riguardante quanto una traccia è ballabile	Float [0, 1]
energy	Energia del brano	Float [0, 1]
instrumentalness	Contenuto relativo di strumenti musicali nella traccia	Float [0, 1]
valence	Metrica riguardante la "positività" della traccia	Intero
liveness	Durata relativa della traccia suonata in una performance dal vivo	Float [0, 1]
loudness	Rumorosità della traccia in decibel (dB)	Float [-60, 0]
release_date	Anno di rilascio del brano	Intero
speechiness	Contenuto relativo di voce umana nella traccia	Float [0, 1]
tempo	BPM della traccia	Float
key	Chiave musicale utilizzata	Factor {0, 1, ..., 11}
mode	Indica se la traccia parte con una progressione armonica	Booleano
explicit	Indica se la traccia è esplicita oppure no (linguaggio volgare)	Booleano
popularity	Popolarità della traccia	Float [0, 100]

2.2.2 Premi

Si noti come la distinzione tra tipo di premio vinto e lo stato in cui è stata ottenuta la certificazione per uno specifico brano, viene fatta solo a scopo dello scraping, in quanto i brani sono così rappresentati sul sito di wikipedia. Tuttavia da questo punto in poi non verrà più tenuto conto di questa informazione, infatti un singolo verrà considerato come **vincitore di un premio** (e quindi di successo) oppure come **non vincitore di un premio** (non di successo).

Il risultato dello scraping della tabella di wikipedia è il seguente:

ATTRIBUTO	DESCRIZIONE	TIPO
title	Titolo della canzone	Stringa
artists	Artisti presenti nella traccia	Stringa
date	Data di rilascio della traccia	Data
genre	Genere musicale del brano	Stringa
award	Premio vinto dal singolo	Stringa {Oro, 1-platino, 2-platino, ...}
nation	Paese in cui è stato vinto il premio	Stringa (Sigla del paese)

2.3 Data integration

Date queste due sorgenti dati, è necessario integrare i dati allo scopo di avere un dataset etichettato, le label saranno appunto se un bravo ha vinto un premio (quindi è di successo) oppure no.

La strategia di entity resolution adottata è considerare un singolo musicale come una singola identità basandosi sul titolo e gli artisti di una canzone. Se il nome del brano musicale è lo stesso e gli artisti coincidono, allora il brano è il medesimo.

2.3.1 Record linkage con MongoDB

A questo scopo i due dataset vengono importati in mongoDB, ogni istanza è rappresentata da un documento. Per la fase di record linkage vengono effettuati i seguenti passi:

1. Import dei dataset in mongodb, inizialmente in due collezioni diverse.
2. Normalizzazione dei dati, i campi di join vengono trasformati in lowercase.
3. Creazione indici.

4. Entrambe le collezioni hanno il campo "artisti" il quale consiste in una stringa contenente la lista degli artisti separati dal carattere ", ". Viene quindi eseguito l'unfold del campo "artista" in entrambe le collezioni, costruendo una lista di per ogni documento facendo uno split sul carattere ", ".
5. Viene eseguita la join sul campo titolo considerando un match valido se l'intersezione tra gli insiemi di artisti dei documenti delle due collezioni non è vuota.
6. Viene ritrasformato il campo artista appiattendo la lista e rappresentando l'insieme degli artisti di un brano come una stringa, separando ogni artista con il carattere ", ".
7. Dump del database in un file .csv, questo è il dataset integrato e verrà usato per il training dei modelli.

2.4 Analisi esplorativa

In questa sezione vengono analizzate e discusse le covariate del dataset.

2.4.1 Assunzioni sul dominio

Anno di uscita singoli

Viene per prima cosa analizzata la distribuzione degli anni di uscita nel dataset.

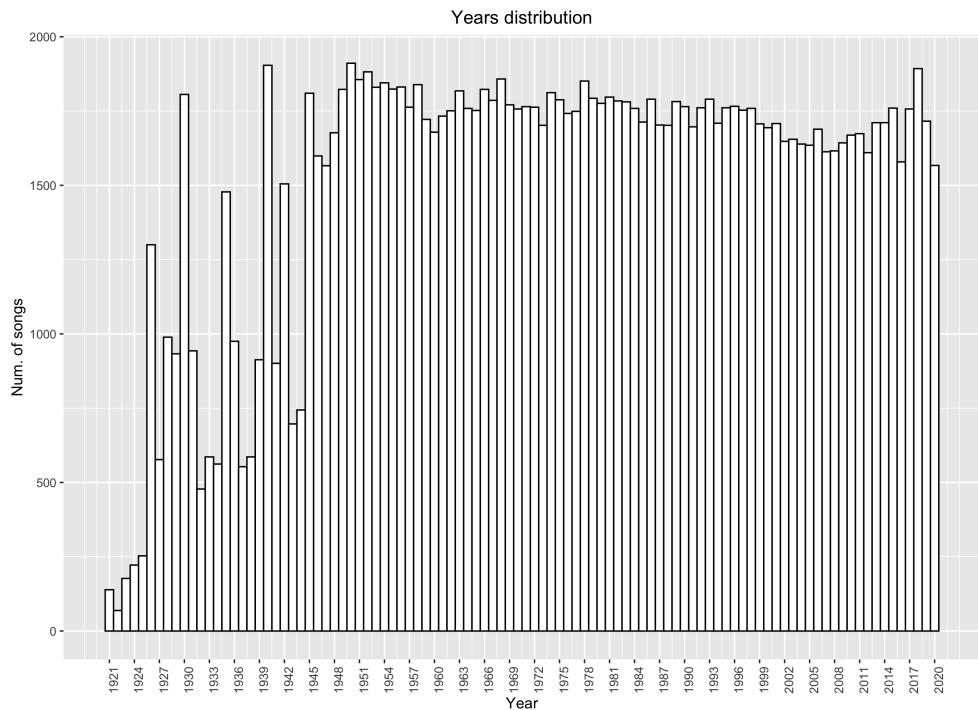


Figure 2.2: Distribuzione anno delle canzoni considerando tutto il dataset.

Dal momento che Spotify è una piattaforma recente, si assume che gli utenti ascoltino maggiormente i brani usciti negli ultimi anni. Inoltre le diverse certificazioni come "Disco d'oro" e "Disco di platino" vengono rilasciate considerando il numero di streaming oltre che alle vendite solo da pochi anni. Inoltre con il passare del tempo i trend musicali cambiano, un fattore determinante per fare diventare una canzone di successo.

Con questa giustificazione si ritiene che sia meglio considerare solo i brani musicali dopo un certo anno di uscita. Prendiamo quindi in considerazione solo le canzoni dopo il 2005.

La distribuzione delle canzoni dal 2005 in poi è rappresentata in Figure 2.2. Viene di seguito mostrato il boxplot delle canzoni dopo quella data distinguendo tra classe positiva e negativa, così da vedere se esistono differenze.

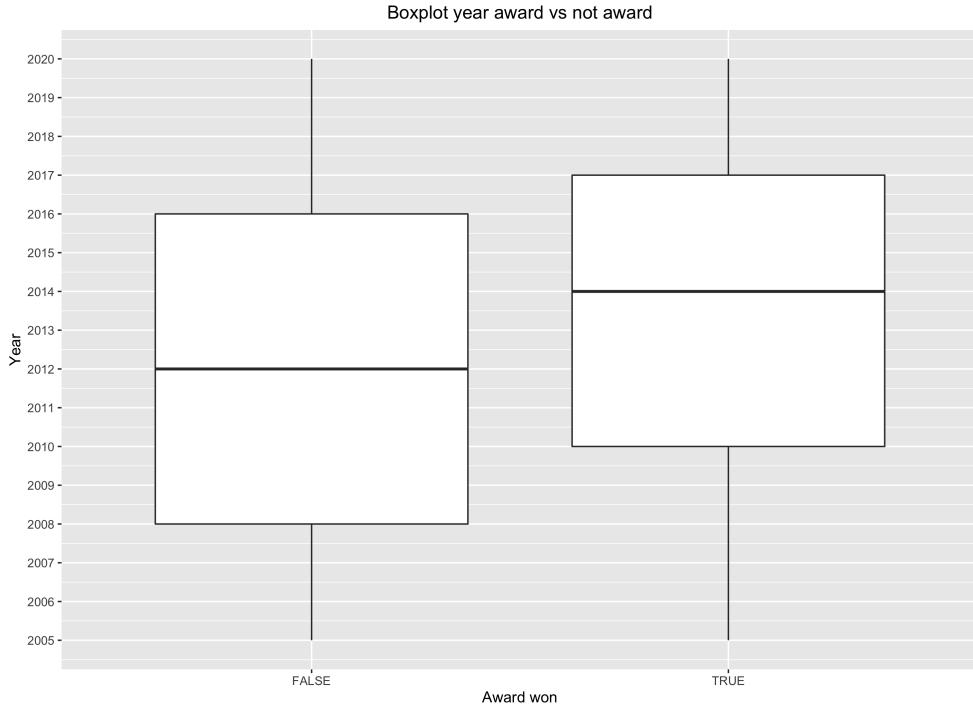


Figure 2.3: Boxplot anno di rilascio, distinguendo tra le classi.

Popolarità

Un altro apsetto che viene analizzato a parte, riguardo il quale è necessario fare ulteriori assunzioni, è il campo popolarità.

Il dataset forintato da Spotify contiene un campo "popularity". Tuttavia come spiegato nella subsection 1.1.2, questo campo non viene utilizzato per etichettare una canzone come di successo, si utilizza invece l'informazione delle certificazioni vinte da una canzone (ottenute da wikipedia).

Il campo "popularity" non verrà usato nella fase di training dei modelli, proprio perché è un dato che non si conosce a priori nel momento in cui un singolo esce, ed è chiaramente influente nel determinare se una canzone vincerà o meno una certificazione e quindi se viene considerata di successo.

Infatti l'informazione sulla popolarità viene calcolata sul numero di streaming dopo un determinato lasso di tempo. Stimare questo valore, dal momento che non è conosciuto fin da subito, è di fatto un problema di regressione ed è molto simile al task di classificazione preso in esame, pertanto il campo viene scartato.

Anche se il campo non viene effettivamente utilizzato, è interessante analizzare la distribuzione dei valori di popolarità delle canzoni nel dataset. Inoltre assumiamo che per il problema trattato, si vogliono classificare delle canzoni che non sono completamente sconosciute. Sarebbe infatti irrealistico pensare che un singolo musicale del tutto sconosciuto

vinca dal nulla una certificazione, risultando come brano di successo.

Si assume di voler utilizzare questi modelli per classificare brani che iniziano ad essere un minimo conosciuti o hanno almeno il potenziale di diventare popolari. Si noti come un brano di poco popolare non implica assolutamente che questo vinca un disco d'oro o di platino.

Viene quindi qui sotto analizzata da distribuzione della popolarità delle canzoni nel dataset.

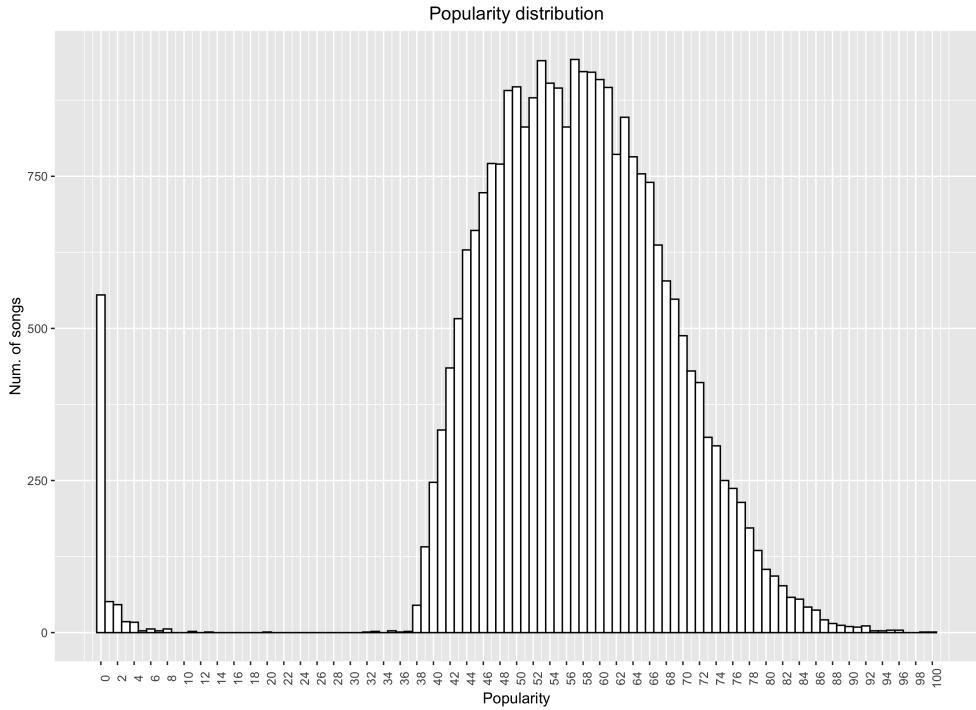


Figure 2.4: Distribuzione popolarità brani musicali.

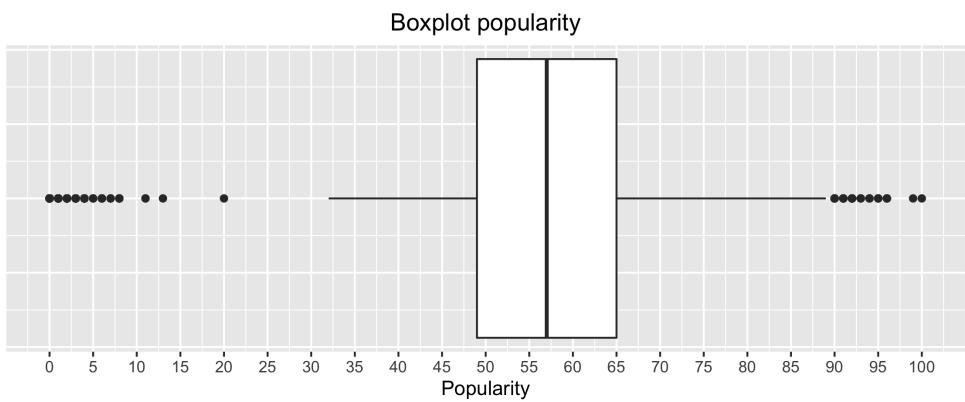


Figure 2.5: Boxplot popolarità dei brani musicali.

Analizzando la distribuzione dei valori della popolarità si considerano il primo secondo e terzo quartile, rispettivamente: $q_{1/4} = 49$; $q_{2/4} = 57$; $q_{3/4} = 65$. Inoltre i valori di media e deviazione standard sono: $\mu = 57.83$; $\sigma = 10.12$.

Analizzando la distribuzione dei valori della popolarità viene scelto 25 come threshold, valore 3 volte più estremo della deviazione standard. Si considerano quindi solo i singoli con il valore "popularity" maggiore del threshold. In questo modo il dataset rispecchia l'assunzione sopra spiegata riguardo la popolarità minima, tuttavia si scartano solo i valori davvero estremi per non rendere questa assunzione troppo restrittiva.

2.4.2 Distribuzione dei valori

Variabili numeriche

Di seguito viene mostrata la distribuzione di tutte le covariate, distinguendo tra singoli che hanno vinto un premio (Classe positiva) e quelli che non hanno vinto un premio (Classe negativa).



Figure 2.6: Pairplot delle covariate.

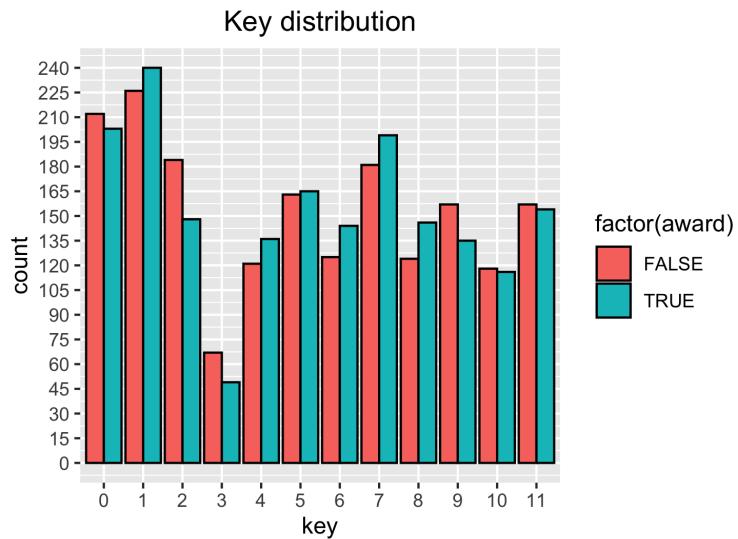
Dal pairplot possiamo notare come non esiste una netta distinzione nei dati tra brani di successo e brani non di successo. Questo sarà sicuramente un problema in fase di classificazione e potrebbe portare a basse performance dei modelli. Riteniamo quindi che sia necessario considerare altre features per ben distinguere brani musicali di successo da quelli non di successo.

Il campo "popularity" ben distingue le due classi, tuttavia come già discusso in section 2.4.1, questa covariata non verrà utilizzata per la creazione dei modelli.

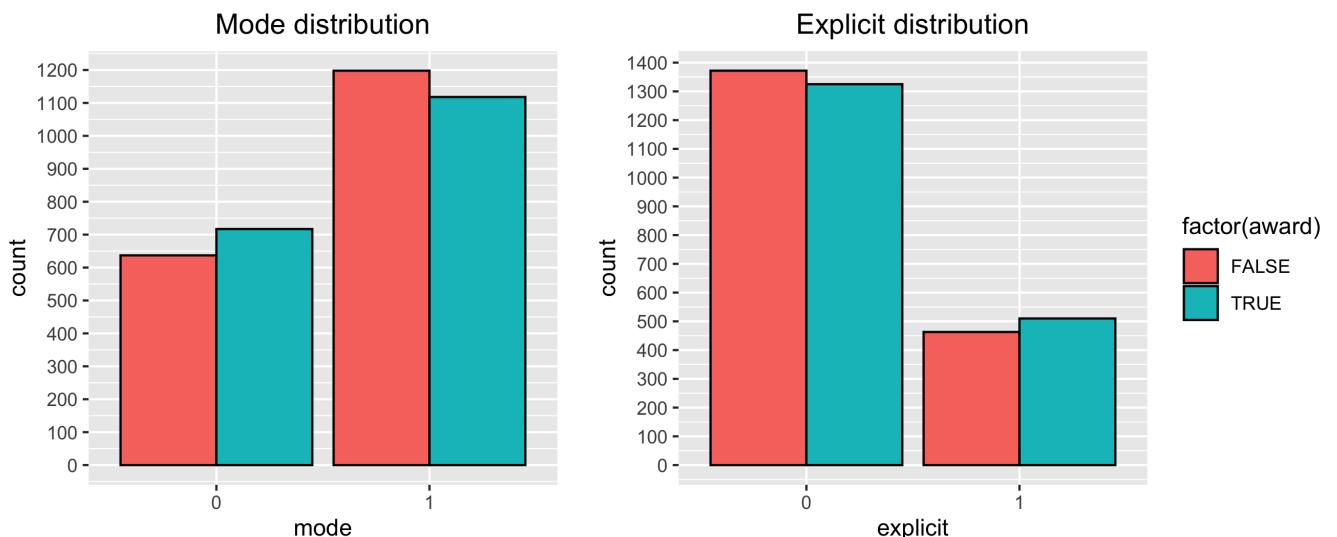
Inoltre esiste correlazione tra alcune covariate, come ad esempio tra "energy" e "loudness", questo rispecchia intuitivamente l'idea di come una canzone più energica è anche più rumorosa. La correlazione tra le variabili viene meglio analizzata in subsection 2.4.4.

Variabili categoriche

Di seguito si esplorano le variabili categoriche del dataset, distinguendo tra classe positiva e negativa.



(a) Variabile key.



(b) Variabile mode.

(c) Variabile explicit.

Figure 2.7: Distribuzione delle variabili categoriche.

2.4.3 Artisti nelle canzoni

Frequenza artisti

Wordcloud

2.4.4 Correlazione features

2.4.5 Principal component analysis

Prima di PCA normalizzazione del dataset

NB: Non facciamo il plot delle prime due componenti principali in quanto varianza spiegata dalle prime due componenti molto bassa

2.4.6 Creazione dataset bilanciato

Undersampling

2.5 Scelta delle features

2.5.1 Variabili scartate

Vengono qui spiegate le features utilizzate

2.5.2 Rappresentazione BOW degli artisti

Artisti in lowercase e rappresentazione bow

Uso del threshold

Threshold frequeunza

Viene usato un threshold a 2 per non avere delle canzoni con artisti completamente sconosciuti. Assumiamo quindi di fare previsioni su canzoni cantante da artisti un minimo conosciuti.

Questa assunzione non è molto restrittiva, ci immaginiamo infatti che per vincere un premio una canzone deve essere cantata da artisti non completamente sconosciuti

Inoltre a causa dell'undersampling è possibile avere diverse canzoni cantate da artisti "sconosciuti" e non avendo scelto una particolare strategia per l'undersampling adottiamo a questo punto l'utilizzo di un threshold

2.5.3 Variabili categoriche

Key - Trasformazione a intero

2.5.4 Coordinate PCA

Dataset proiettato componenti principali

2.5.5 Normalizzazione e standardizzazione

Conversione in lowecase Converto in variabili numeriche e factor + label corretta

2.5.6 Dataset proiettato componenti principali

2.5.7 Riassunto feature finali utilizzate

Tabella

Chapter 3

Campagna sperimentale

3.1 Approccio

3.1.1 10-folds cross validation

3.1.2 Training set e test set

3.1.3 Model selection

Ottimizzazione iperparametri

Grid search

3.2 Misure di performance

3.2.1 Accuracy

3.2.2 Precision, Recall e F-measure

3.2.3 Curve ROC

3.3 Support Vector Machine

3.3.1 Kernel

3.4 Decision Tree

3.5 Esperimenti

3.5.1 Performance

3.5.2 Modelli a confronto

Chapter 4

Conclusioni

Perchè le performance sono basee?

Spotify genera le caratteristiche con un algoritmo quindi sicuramente un po' approssimato

E' effettivamente difficile capire se una canzone vincerà un premio, dipende da molti fattori