

# Selección automática de la siguiente canción, basada en las listas de reproducción del servicio de música en línea: Spotify.

Javier Eduardo Jaimes Velasquez<sup>1</sup>, Paula Nathalia Pineda Ortiz<sup>2</sup>, Brandon Valencia Murillo<sup>3</sup>, Johan Alejandro Cifuentes Gonzalez<sup>4</sup>, Melvin Damar Pineda Cañon<sup>5</sup>

2024-06-07

## Resumen

Las plataformas de música en línea deben emplear una forma eficiente y eficaz que resuelva uno de los problemas más críticos, recomendación de música automática. En este documento se explora uno de tantos métodos expuestos en la actualidad para resolver la tarea. Se espera poder evaluar su funcionamiento respecto a otros y también su efectividad a la hora de resolver el problema.

---

<sup>1</sup> Politécnico Grancolombiano, [jajames4@poligran.edu.co](mailto:jajames4@poligran.edu.co)

<sup>2</sup> Politécnico Grancolombiano, [papineda1@poligran.edu.co](mailto:papineda1@poligran.edu.co)

<sup>3</sup> Politécnico Grancolombiano, [bvalencia6@poligran.edu.co](mailto:bvalencia6@poligran.edu.co)

<sup>4</sup> Politécnico Grancolombiano, [jcifuntes@poligran.edu.co](mailto:jcifuntes@poligran.edu.co)

<sup>5</sup> Politécnico Grancolombiano, [mdpineda1@poligran.edu.co](mailto:mdpineda1@poligran.edu.co)

## 1. Introducción

Las canciones se volvieron parte de cotidianidad, y por ellos los servicios de música en línea juegan un papel importante en la vida de cada ciudadano. Los consumidores de música ya no necesitan un dispositivo de propósito dedicado, como el radio. Ahora la música está disponible en una computadora, o en las manos a través de un smartphone.

Las plataformas de música en línea abundan, de acuerdo con “*Wikipedia List of Online Music Databases*” (*n.d.*), la variedad es extensa. Por ende, cada proveedor debe esforzarse en crear una experiencia única en un mercado abundado de opciones.

Dichas plataformas están compuestas de muchas o tal vez cientos de funcionalidades, pero una en ella es clave: la recomendación automática de canciones. De ella depende la permanencia de muchos de sus clientes. La capacidad intrínseca del sistema de proveer de forma automática la siguiente canción es clave a la hora de hacer clientes felices. Como se menciona en *Pichl, Zangerle, and Specht (2016)*: “*Un entendimiento profundo de los atributos de las listas de reproducción y como los usuarios crean y mantienen dichas listas puede contribuir de manera natural a mejores recomendaciones y personalizaciones*”.

En este documento se espera poder responder la pregunta: ¿cuál es el método de selección de música automática que mejor contribuye a una mejor experiencia de los usuarios de la plataforma de música en la línea con foco en Spotify?

En el artículo *Lu et al. (2016)*, se menciona diferentes enfoques que resuelven nuestra pregunta. Para el desarrollo de la pregunta en este documento se decidió elegir el enfoque basado en contenido. Si bien este método no es el más usado en sistemas de recomendación se espera que al menos pueda plantear una forma eficaz de resolver la pregunta.

## 2. Análisis Exploratorio de Datos

A continuación, se describe la metodología usada en el análisis de los datos.

### 2.1. Conjunto de Datos

Se obtuvo acceso al conjunto de datos [Million Playlist Dataset](#) publicado por **Spotify** en la plataforma de [AI Crowd](#). A partir de la fuente original tomamos como referencia una muestra de 1000, en las que aleatoriamente se escogieron 100.

Para una idea inicial del conjunto de base se encuentra: *datasets/spotify\_millions\_playlist/mpd.slice.0-999.json*.

```
library(jsonlite)
library(dplyr)
```

```

library(magrittr)
library(jsonlite)
library(tidyr)
library(spotiflyr)

Sys.setenv(SPOTIFY_CLIENT_ID = "XYZ")
Sys.setenv(SPOTIFY_CLIENT_SECRET = "XYZ")

# Archivo original de Spotify
t <- fromJSON("datasets/spotify_millions_playlist/mpd.slice.0-999.json")

# Lista de Reproducciones
d <- t$playlists %>% select(c("name", "collaborative", "pid", "modified_at", "num_tracks", "num_albums", "num_followers", "tracks"))

# Se toma como muestra 100 lista de reproducciones.
set.seed(100)
srow <- sample(1:nrow(d), 100)
p <- d[srow, ]
tr <- do.call("rbind", p$tracks) %>% select(-c("pos")) %>% distinct()

# Obtención de datos a través del API de Spotify
access_token <- get_spotify_access_token()

spotify_track_uris <- as.list(tr$track_uri)
spotify_tracks_ids <- lapply(spotify_track_uris, function(x) strsplit(x, ":")[[1]][3])
spotify_tracks_id_by_100 <- split(spotify_tracks_ids, ceiling(seq_along(spotify_tracks_ids) / 100))

audio_features_by_ids <- get_track_audio_features(gsub(" ", "", toString(spotify_tracks_id_by_100[1][[1]])))

for (i in 2:length(spotify_tracks_id_by_100)) {
  audio_features_by_ids <- rbind(audio_features_by_ids, get_track_audio_features(gsub(" ", "", toString(spotify_tracks_id_by_100[i][[1]]))))
  Sys.sleep(1)
}

trf <- left_join(x = tr, y = audio_features_by_ids, by = c("track_uri" = "uri")) %>% na.omit()

pj <- toJSON(p, auto_unbox = TRUE)
write_json(pj, path = "data/playlist.json")

pf <- p %>% select(-c("tracks"))

```

```
write.csv(pf, "data/playlists.csv", row.names = FALSE, quote = FALSE)
write.csv(trf, "data/tracks.csv", row.names = FALSE, quote = FALSE)
```

A partir de la muestra se construyen los siguientes conjuntos de datos, con información obtenida a través del [API de Spotify](#).

## 2.2. Estructura del Conjunto de Datos

Se describen los datos y sus tipos con una breve descripción.

### 2.2.1. Lista de Reproducciones

La **lista de reproducción** tiene la siguiente estructura:

Nombre del Campo	Tipo	Descripción
<b>name</b>	Caracteres	Nombre de la lista
<b>Collaborative</b>	Lógico	Pública o Privada
<b>pid</b>	Entero	Identificador Único
<b>modified_at</b>	Entero	Tiempo de modificación
<b>num_tracks</b>	Entero	Número de canciones
<b>num_albums</b>	Entero	Número de álbumes
<b>num_followers</b>	Entero	Número de seguidores

### 2.2.2. Canciones

Se debe aclarar que, cada lista de reproducción contiene asociado un numero  $n$  de canciones con la

siguiente estructura.

Nombre del Campo	Tipo	Descripción
<b>artist_name</b>	Caracteres	Nombre del artista
<b>track_uri</b>	Caracteres	Identificador Único de Canción
<b>artist_uri</b>	Caracteres	Identificador Único de Artista
<b>track_name</b>	Caracteres	Nombre de Canción
<b>album_uri</b>	Caracteres	Identificador Único de Álbum
<b>duration_ms.x</b>	Real	Duración en ms
<b>album_name</b>	Caracteres	Nombre de Álbum
<b>danceability</b>	Real	Es bailable
<b>energy</b>	Real	Es activa y enérgica
<b>loudness</b>	Real	Tienen voces
<b>speechiness</b>	Real	Sin voces
<b>acousticness</b>	Real	Es versión acústica
<b>instrumentalness</b>	Real	Es versión instrumental
<b>liveness</b>	Real	Esta en vivo

---

tempo	Real	BPM

---

## 2.3. Estadísticas descriptivas

A continuación, se presentan los primeros resultados de nuestro análisis exploratorio inicial.

### 2.3.1. Resumen Lista de Reproducciones.

```
library(jsonlite)
library(dplyr)
library(magrittr)
library(jsonlite)
library(tidyr)
library(flextable)

playlists <- read.csv("data/playlists.csv")

# Se clona la lista de canciones (playlist) para propósitos de visualización
data_summary <- playlists[, c("collaborative", "num_tracks", "num_albums", "num_followers")]
data_summary$collaborative <- ifelse(data_summary$collaborative, "Public", "Private")
summary <- data_summary %>% summarizer(by = c("collaborative"), overall_label = "Total")
summary_table <- summary %>% as_flextable(spread_first_col = TRUE)
summary_table
```

	Private (N=97)	Public (N=3)	Total (N=100)
num_tracks			
Mean (SD)	69.3 (59.7)	27.7 (20.4)	68.1 (59.3)
Median (IQR)	52.0 (74.0)	23.0 (20.0)	49.5 (70.2)
Range	6.0 - 238.0	10.0 - 50.0	6.0 - 238.0
num_albums			
Mean (SD)	50.8 (43.3)	24.3 (16.3)	50.0 (42.9)

	Private (N=97)	Public (N=3)	Total (N=100)
Median (IQR)	41.0 (51.0)	21.0 (16.0)	39.5 (51.0)
Range	3.0 - 194.0	10.0 - 42.0	3.0 - 194.0
num_followers			
Mean (SD)	1.5 (1.5)	1.0 (0.0)	1.5 (1.5)
Median (IQR)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
Range	1.0 - 11.0	1.0 - 1.0	1.0 - 11.0

### 2.3.2. Distribucion de Listas de Reproducción

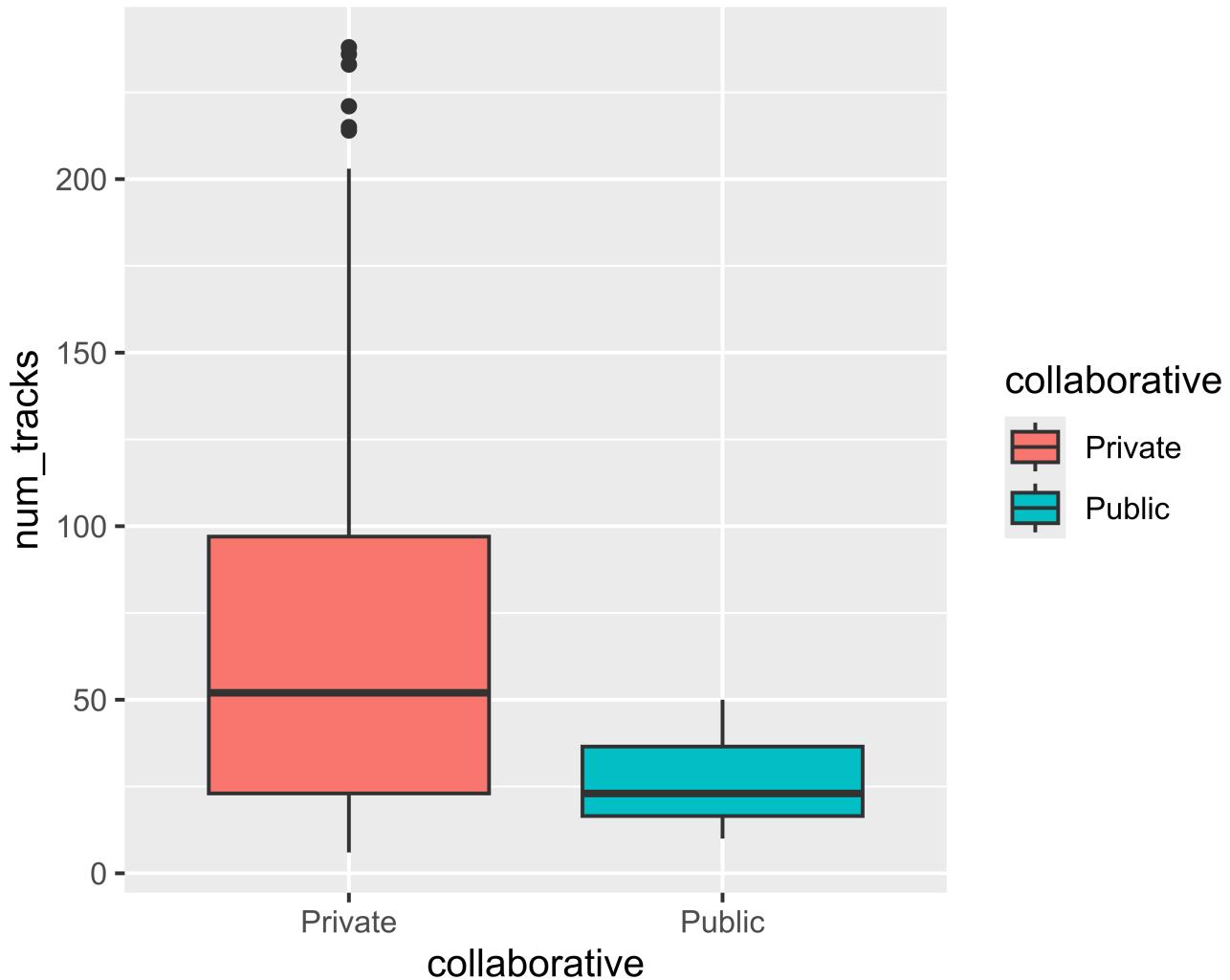
```

library(tidyverse)
library(hrbrthemes)
library(viridis)
library(magrittr)
library(ggplot2)
library(dplyr)
library(flextable)

playlists <- read.csv("data/playlists.csv")

playlists_by_group <- playlists %>% group_by(collaborative) %>% summarise(mnum_tracks=mean(num_tracks))
playlists %>% ggplot( aes(x=collaborative, y=num_tracks, fill=collaborative)) + geom_boxplot()

```

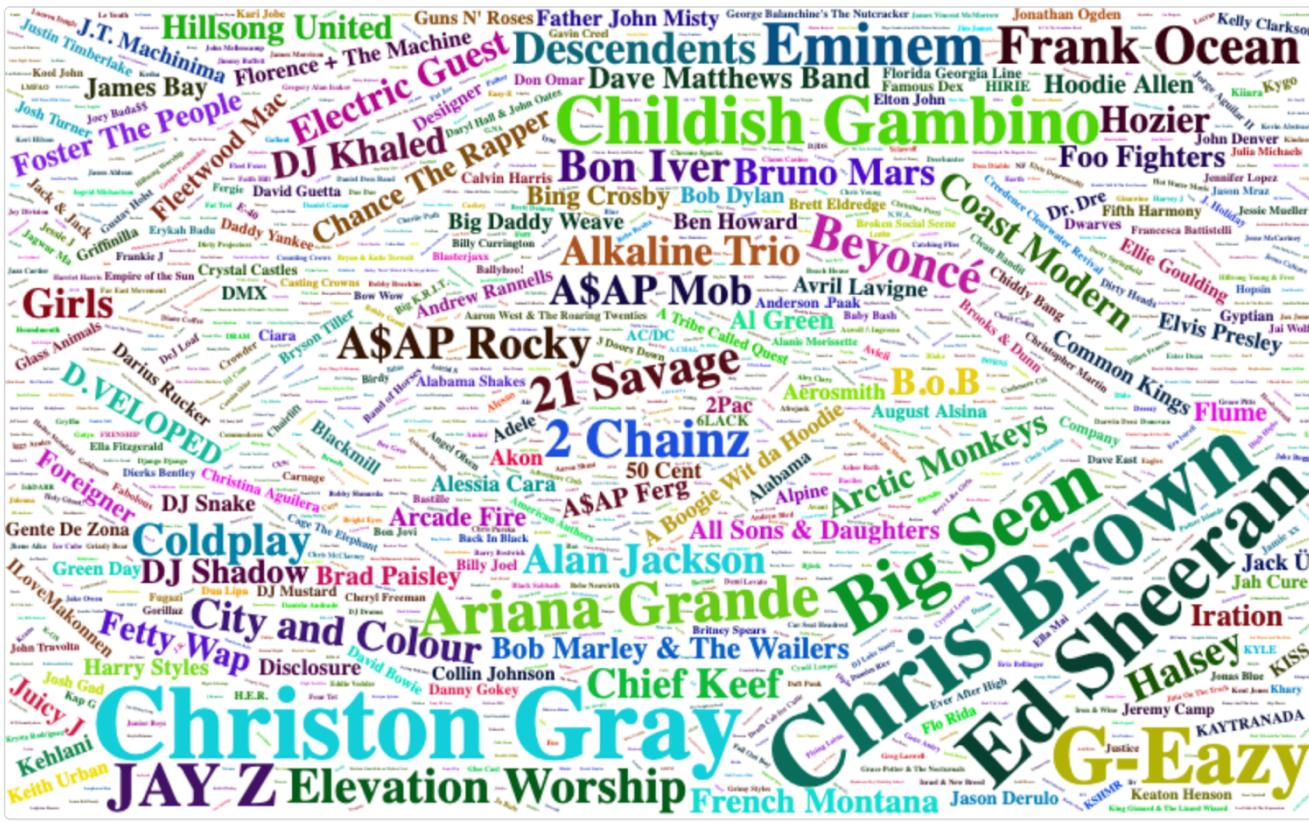


### 2.3.3. El artista con más reproducciones

```
library(wordcloud2)
library(readr)
library(magrittr)
library(dplyr)

dtracks <- read_csv("data/tracks.csv")
wc_tracks <- dtracks %>% group_by(artist_name) %>% summarise(n = n_distinct(track_uri))

wordcloud2(data = wc_tracks, size = 2)
```



### 2.3.4. La canción con más reproducciones

```
library(wordcloud2)
library(readr)
library(magrittr)
library(dplyr)

dtracks <- read_csv("data/tracks.csv")
wc_tracks <- dtracks %>% group_by(album_name) %>% summarise(n = n_distinct(track_uri))

wordcloud2(data = wc_tracks, size = 2)
```



### 3. Conclusiones

Se puede apreciar información relevante sobre el análisis exploratorio. La mayoría de las listas de reproducción son privadas, en comparación con las listas de reproducción pública. Esto refuerza la idea inicial de que al crear un modelo basado en listas de reproducción (Contenido generado por usuario) podría llevar a aumentar la satisfacción en la experiencia de los usuarios de las plataformas de música en línea. Los artistas y álbumes más escuchados podrían utilizarse como características del modelo de recomendación de música automática. En buena medida reflejan la relación que existe entre los usuarios dueños de sus listas de reproducción y sus preferencias musicales. Vale la pena mencionar que uno de los artistas con más reproducciones es Chris Brown y el álbum con más reproducciones es 24K Magic.

#### 4. Referencias Bibliográficas

- Pichl, M., E. Zangerle, and G. Specht. 2016. "Understanding Playlist Creation on Music Streaming Platforms." In *2016 IEEE International Symposium on Multimedia (ISM)*, 475–80. Los Alamitos, CA, USA: IEEE Computer Society. <https://doi.org/10.1109/ISM.2016.7770030>.

IEEE Computer Society. <https://doi.org/10.1109/ISM.2016.0107>.

“Wikipedia List of Online Music Databases.” n.d.  
[https://en.wikipedia.org/wiki/List\\_of\\_online\\_music\\_databases](https://en.wikipedia.org/wiki/List_of_online_music_databases).