# Prueba técnica - Jr. Data Scientist

# GUROS

**José Armando López Velasco**

https://www.linkedin.com/in/jos%C3%A9-armando-l%C3%B3pez-velasco-a61092167/

## 1 Introducción

Los requerimientos de este problema se podrán ver en la carpeta "docs" ubicada en este repositorio[https://github.com/ArmandoLp/GUROS].

El problema central consiste en resolver la pregunta: **¿Cuáles son los 5 artistas más prometedores del Q3 2021?**

Primero debemos definir que significa "prometedores" dada la definición:

Prometedor:[persona, cosa] Que promete o da muestras de que va a triunfar o a resultar bueno en el futuro.

Podemos decir qué podemos descartar como característica de prometedor la cantidad de followers o popularidad de manera directa, pero si serán útiles para fijar a un artista prometedor, por lo que en este documento se hará un acercamiento general utilizando la información sumamente limitada que proveé la API de Spotify.

#### ¿A qué se refiere "Q3 2021"?

Al periodo contenido entre los meses 7 y 9 (incluyendo los extremos) del año 2021, por lo que una parte fundamental es limitar la búsqueda a este periodo

## 2 ¿Cómo resolveré el problema?

La idea inicial que tenía para abordar el problema era identificar la cantidad de reproducciones de las canciones dada la ventana de tiempo, con base en eso fabricar una variable que tuviese en cuenta la cantidad de followers que un artista tiene, de esta manera marginar la ventaja de artistas conocidos, dividiendo la cantidad de reproducciones entre el total de followers, de esta manera, el que tuviese el cociente de mayor tamaño indicaba que tuvo un éxito inesperado en una particular canción.

## 2.1 Problemas encontrados

Me di cuenta que la API de Spotify está muy orientada a su manipulación, más no a la explotación de datos, por lo que la idea central de cómo resolverlo no tenía mucho sentido. Bajo estas limitaciones me planteé el siguiente flujo de objetivos:

1. Identificar Playlist del usuario spotify que tengan a los artistas más relevantes
2. Extraer todas las canciones
3. Extraer todos los artistas
4. Filtrar por fecha de lanzamiento, deben estar dentro del Q3
5. Con base en distintas métricas disponibles (y algunas métricas generadas) ordenar el top 5 de artistas más prometedores.

### 2.1.1 Identificar Playlist del usuario spotify que tengan a los artistas más relevantes

Obtenemos todas las playlists del usuario 'spotify'

Dada la limitación, sólo podemos obtener 50 playlist, podríamos hacer esta consulta cíclicamente, pero para este ejercicio lo haremos sólo una vez y agregaremos la playlist “Mega Éxitos 2021”.

### 2.1.2 Extraer todas las canciones

Una vez identificadas las playlist, extraemos todos los id de las canciones que las conforman, obteniendo un total de 3894 canciones, de las cuales 3339 son canciones únicas.

#### 2.1.2.1 Filtrar por fecha

Obteniendo la información por canción, podemos identificar la fecha.

Definimos la ventana para el Q3, a tomar en cuenta que el inicio debe ser incluido (>=) y el final no (<)

De esta manera, una vez filtrado por el atributo del json ‘release_date’ podemos delimitar el problema con todas las canciones publicadas durante el Q3 2021.

# 2.1.3-4 Extraer todos los artistas

Una vez que tenemos todas las canciones, podemos conocer a los artistas que participaron, la idea de hacer esta extracción es generar dos datasets principales:

a) Contiene toda la información de todas las canciones

b) Contiene toda la información de los artistas

Con estos dos se planea generar un tercer dataset que será la mezcla de ambos, esto con la finalidad de poder hacer análisis más rápido, debido al tamaño de cada uno, esta mezcla no representa gran problema.

### 2.1.5 Generamos el dataset final para la extracción de métricas

Una vez generados los dos datasets, procedemos a construir el tercero con un simple merge, con la finalidad de tener toda la información relacionada y muy descriptiva en un sólo lugar, esto facilitará los querys futuros y la generación de métricas específicas.

# 3 Presentación de resultados

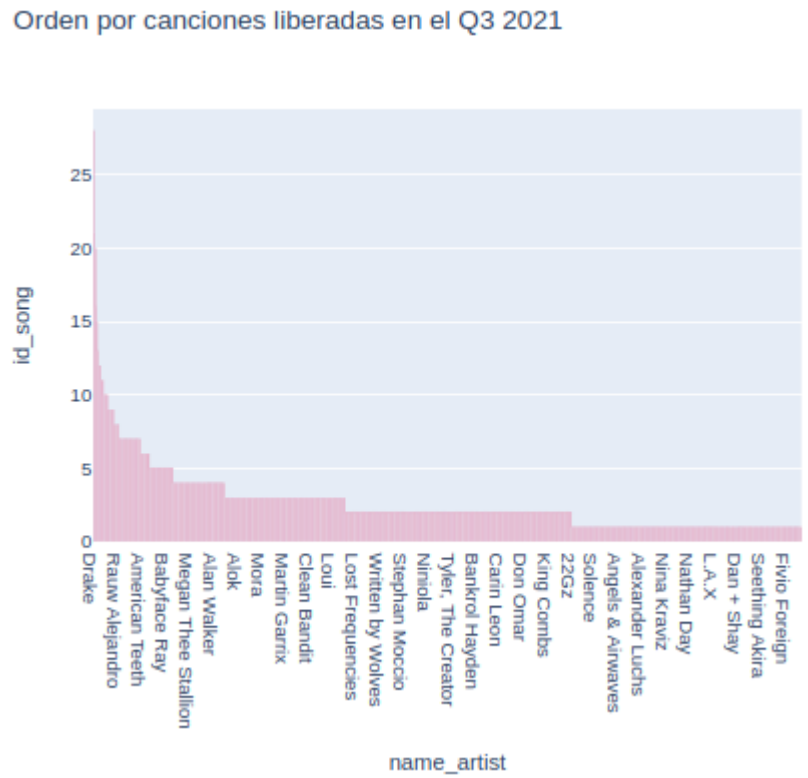
El primer acercamiento que haremos consiste en ordenarlos por la cantidad de apariciones que tuvieron en el periodo. ¿Por qué? podríamos considerar como artista prometedor al artista que se encuentra trabajando constantemente, liberando nuevas canciones y colaboraciones, sin embargo, no creo que este sea el criterio adecuado debido a que los artistas más populares deberían tener más incidencias por el simple hecho de ser populares. De cualquier manera se aborda la búsqueda de la solución por esta vía para agotar todas las posibles presuposiciones.

El top 5 resultante de este orden es:

```
Drake            28
Lil Durk         21
Polo G           20
Ty Dolla $ign    16
G-Eazy           15
Young Thug       13
Lil Baby         12
DJ Drama         12
Lakeyah          12
Lil Wayne        11
```

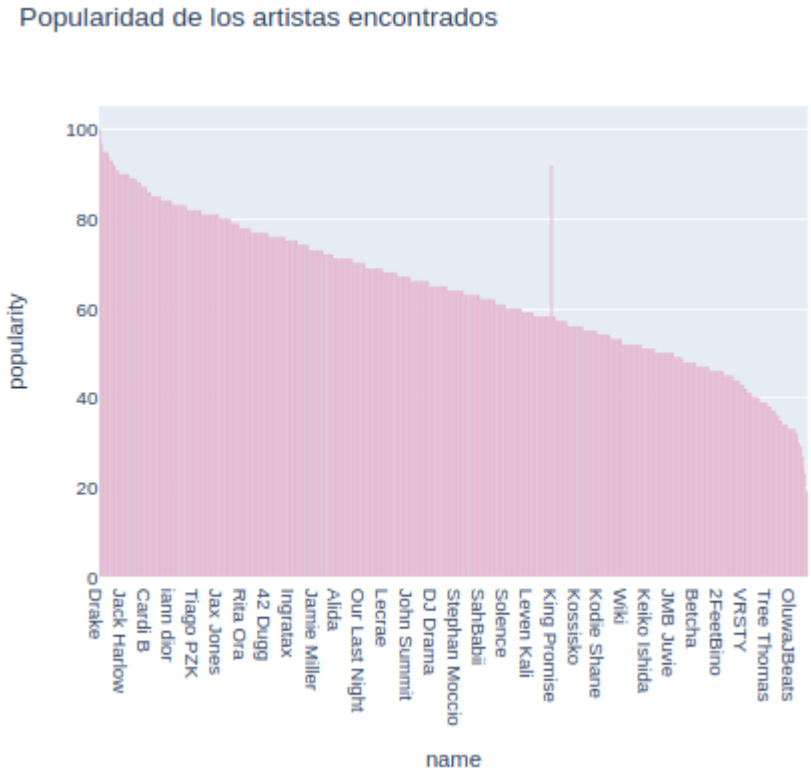
Como era de esperarse, los artistas más conocidos aparecen aquí, con una cantidad inmensa de canciones para un periodo de 3 meses.

La distribución general, se ve de la siguiente manera:

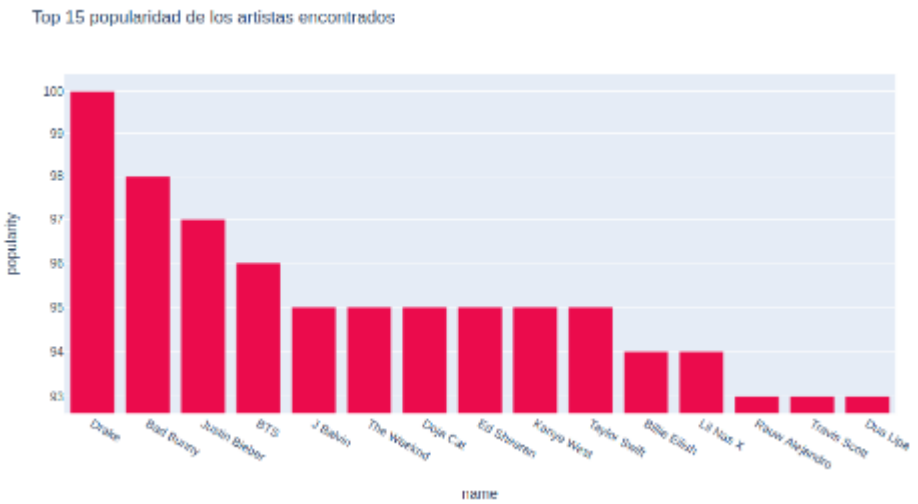



Podemos notar que Drake es el artista que más canciones liberó durante el Q3, sin embargo esto no implica que sea el más prometedor.

Otro parámetro interesante es la popularidad, veamos la siguiente gráfica de su distribución:




El top 15 sería:

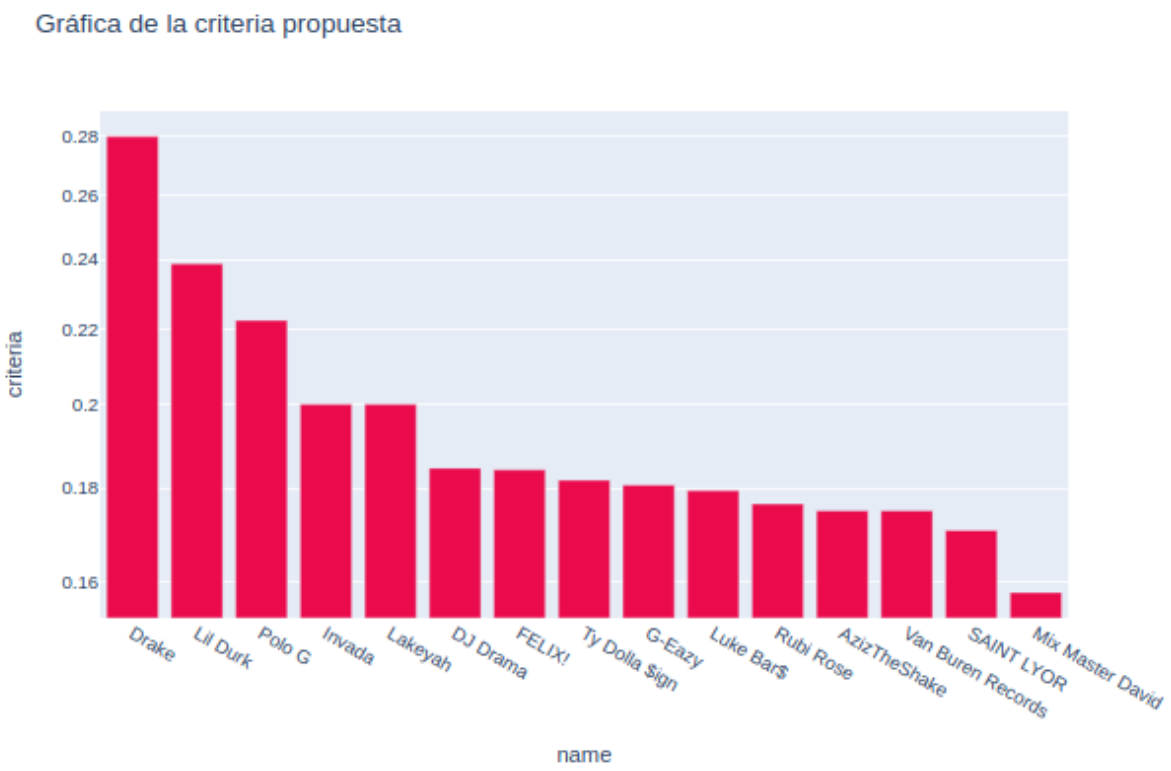



Podemos notar que Drake liberó más canciones y además es el más popular. Lo cuál podría ser un muy buen indicador, sin embargo, artistas poco seguidos pueden no aparecer, mientras que su carrera podría ser prometedora.

# ¿Lo estamos haciendo correctamente?

Sabemos que la popularidad puede ayudar a un artista a producir más canciones, por lo que es una competencia desbalanceada, por esta razón se propone incluir un criterio propio, el cuál tome en consideración la cantidad de canciones dada la popularidad que posee un artista

El criterio definido es:
$criteria = A_{count-songs}/A_{popularity}$
Donde: $A$: $Artista$
Es decir, marginamos la cantidad de producciones dada la popularidad, de esta manera, relajamos un poco esa ventaja. El resultado es:




Drake, bajo este criterio sigue estando en el top 1.

Nuestro criterio penaliza más a un artista popular que no ha tenido colaboraciones, pero premia más a un artista poco popular con una cantidad considerable de contribuciones durante el Q3. Notamos los cambios a partir del p2

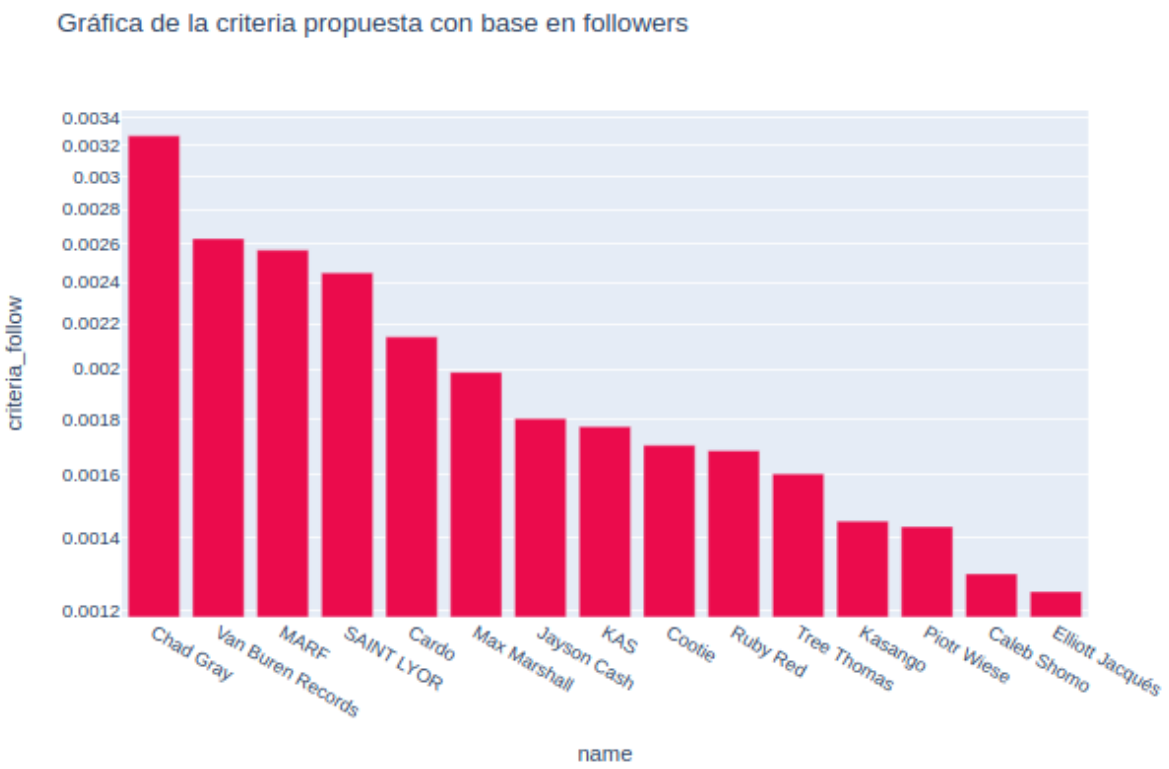
### ¿Cómo se verá este criterio con la cantidad de followers?

# ¿Qué problema tenemos?

Que una cantidad extremadamente pequeña de followers hará que el artista aparezca con un score muy alto, por lo que eliminaremos a todos los que tengan menos de 1000 followers

De esta forma tenemos:
$criteria_{follow} = A_{count-songs}/A_{followers}$

De la misma manera, tenemos una marginación parecida, pero esta vez delimitada por la cantidad de followers. Dando como resultado:




Ahora se encuentra en top1 Chad Gray, lo cual es una gran noticia, ya que para este criterio, la popularidad de Drake no ha afectado a la salida. Sin embargo esto sucede porque este criterio castiga demasiado a los artistas con muchos followers, y premia demás a los artistas con pocos followers, los cuales con pocas apariciones podrían ser considerados como promesas erróneamente, por lo que la solución propuesta debe considerar ambos criterios.

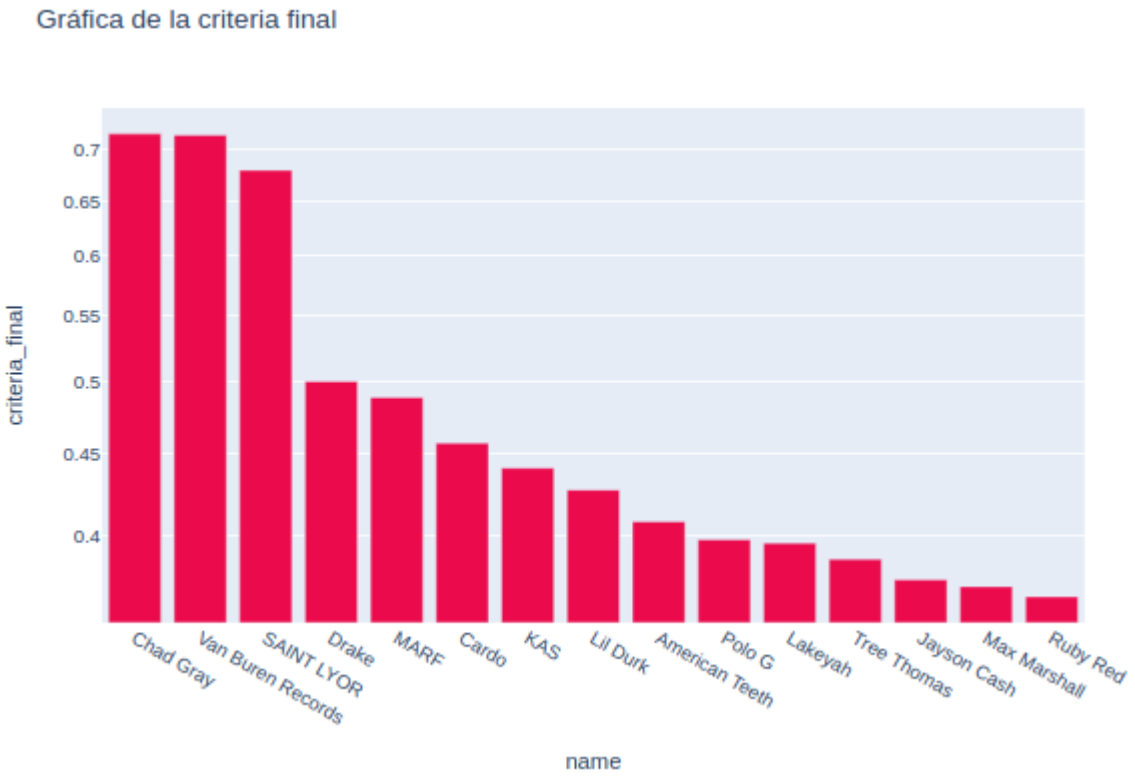
De esta manera definimos el criterio final

# Criterio final

Este criterio tratará de ponderar ambos acercamientos, el cual beneficia a los artistas poco conocidos con pocas canciones exitosas, y el criterio que premia la popularidad, de esta manera tenemos un criterio balanceado, el cual se define:

$$criteria_{final} = (criteria_{follow}/max(criteria_{follow}) \quad + \quad criteria \,/\, max(criteria))/2$$

Calculando este criterio, tenemos:




# ¿Nuestro criterio funciona?

La correlación entre el criterio y las otras dos variables debería ser bajo, esto debido a que a pesar de componerse de ellos, el resultado debe estar más apegado a un factor sin preferencias por popularidad o followers, la correlación resultantes es:

| | followers.total | criteria_final | popularity |
|---|---|---|---|
| **followers.total** | 1.000000 | -0.021199 | 0.511588 |
| **criteria_final** | -0.021199 | 1.000000 | -0.190498 |
| **popularity** | 0.511588 | -0.190498 | 1.000000 |

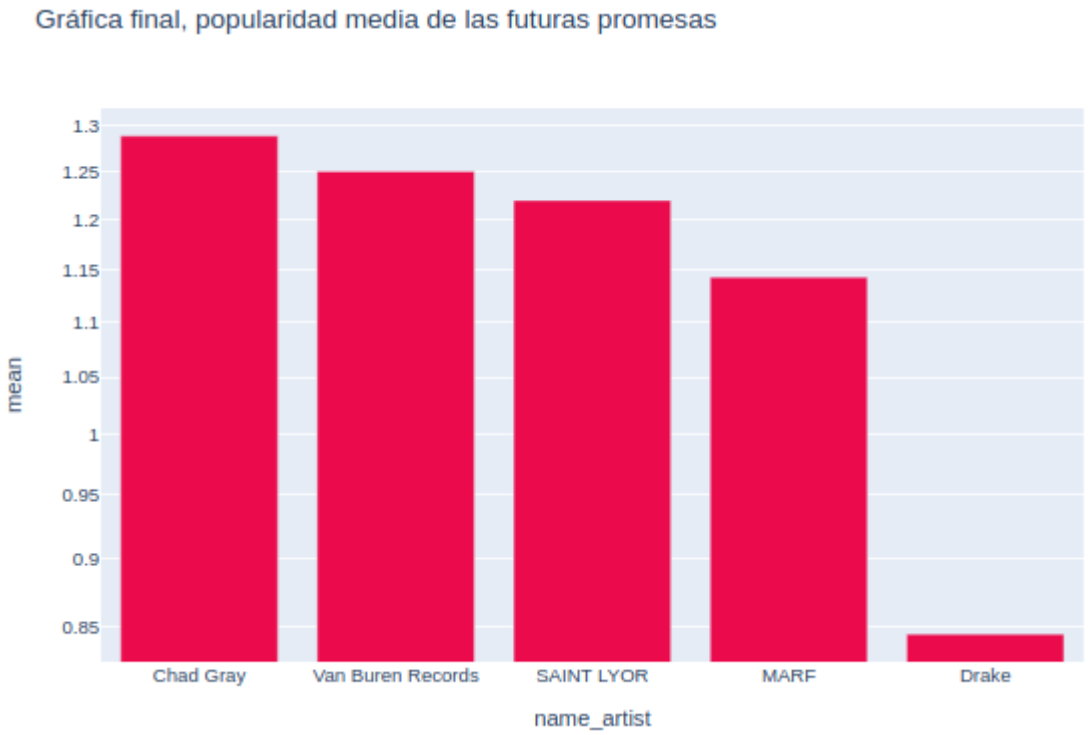
Lo que valida que tenemos razón.

# Análisis del top 5 obtenido por nuestro criterio

Al calcular la media y desviación estándar de la popularidad de sus canciones nos podemos dar cuenta de su promesa futura, esto, porque podemos asumir que la media de la popularidad dado el orden de nuestro criterio, garantiza que el éxito de un artista es constante.

| | name_artist | mean | std |
|---|---|---|---|
| **0** | Chad Gray | 1.287879 | 0.052486 |
| **4** | Van Buren Records | 1.250000 | 0.000000 |
| **3** | SAINT LYOR | 1.219512 | 0.000000 |
| **2** | MARF | 1.142857 | 0.000000 |
| **1** | Drake | 0.844643 | 0.066638 |

Así, obtenemos el resultado final:




# Conclusión

Dada la ambigüedad del problema, la solución podría tener múltiples respuestas "correctas" sin embargo, traté de agotar todos los posibles caminos, atenuando el factor inevitable donde la popularidad rigiera el top 5, por lo que podemos decir que el criterio final es bastante razonable y sobre todo justo, midiendo los 5 artistas (sin importar seguidores) que más se encuentran en un camino hacia el éxito.

# Trabajo futuro

Sin las limitaciones sobre el uso forzoso de la api, hubiese extraído los datos de:https://spotifycharts.com/regional/global/daily/latest Extrayendo por día el top 50 de canciones más escuchadas, filtrando los artistas y hacer un criterio similar al propuesto en este trabajo.

Al tener la marca temporal se hubiese podido generar gráficas con líneas con base al tiempo, de esta manera podríamos ver si nuestro criterio está siendo alto aún cuando el artista está bajando su performance en las listas.

Otro posible acercamiento sería utilizar el dataset que se mencionó en el artículo, replicar su modelo de ML y generar la predicción para estos artistas, ordenarlo conforme al criterio y la probabilidad promedio de éxito, y así obtener un panorama más amplio sobre "prometedores" ya que la misma palabra hace referencia al "futuro" por lo que me parece que esta propuesta sería la más adecuada. Pueden checar mi github para ver modelos similares de clasificación, sólo como precedente para considerar esta solución como realizable.