

# GÉNEROS MUSICALES

EN

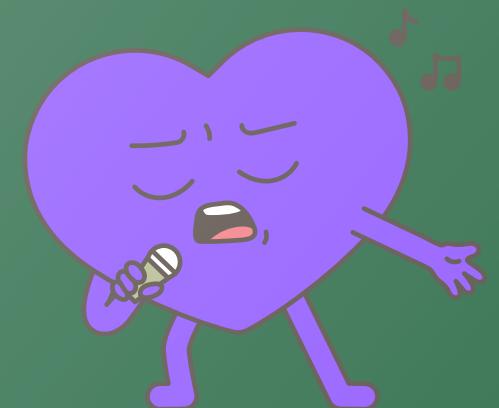


Spotify®

**Joan Espada  
Alejo Razeto  
Ezequiel Coggiola**

INTRODUCCIÓN AL APRENDIZAJE AUTOMÁTICO

# ¿Se puede predecir el género musical de una canción?



# DATASET

## Music Genre Classification

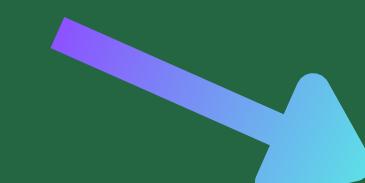
**Fuente:** Kaggle

Contiene **25709** canciones (filas) tomadas de **Spotify**

Con **16** atributos (columnas) por canción:

- Popularity
- danceability
- energy
- key
- loudness
- mode
- speechiness
- acousticness
- instrumentalness
- liveness
- valence
- tempo
- duration\_in\_min/ms
- time\_signature
- Class

Tomamos los atributos que tengan la capacidad de describir un género musical



Tienen valores que van entre 0 y 1

Nuestro target va a ser 'Class'

El género de la canción

# LIMPIEZA

## Eliminación de NaN's:

Class (7.713)

Instrumentalness (6.286)

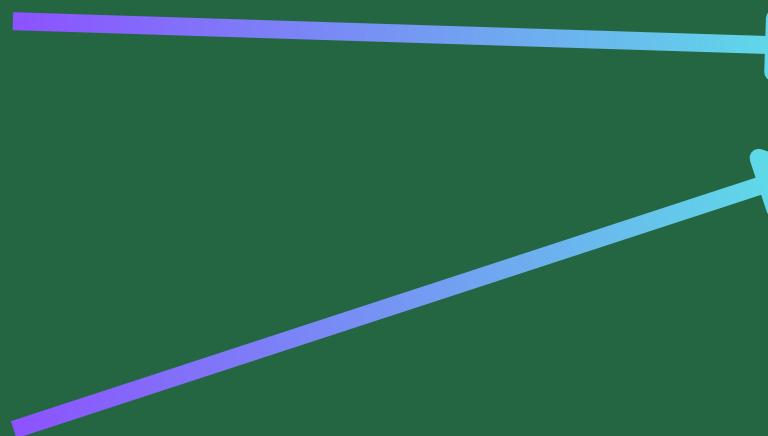
Popularity (655)

Eliminación de duplicados: 325

## Géneros a predecir

- **Rock:** 3876
- **Indie Alternativo:** 2330
- **Metal:** 1697
- **Pop:** 1263
- **Música Alternativa:** 1198
- **Blues:** 1130

- **Hip-Hop:** 578
- **Instrumental:** 575
- **Folk Acústico:** 487
- **Bollywood:** 305
- **Country:** 180



**Cantidad final**  
13.306 canciones

## Train-Test

% de datos en train = 70 %  
% de datos en test = 30 %

# METRICAS USADAS

**¿Que métricas utilizamos para medir el modelo?**

**Accuracy:** Nos da una certeza de si los datos fueron correctamente clasificados.

**Recall:** Para ver que los datos que fueron en sus respectivos géneros estén bien clasificados.

**f1-score:** Promedio de estas dos.

# MODELO

## ¿Por que RFC?

Tenemos un problema de clasificación con múltiples clases. RandomForest es considerado un buen modelo para estos casos, con buena generalización y poco sobreajuste.

### RandomForest:

```
RFC = RandomForestClassifier(random_state = 42, n_jobs=-1)
```

### Validación Cruzada:

```
cross_validate(RFC, X_train , y_train, scoring = "accuracy", n_jobs = -1, return_estimator = True)
```

## ¿Qué es el mejor modelo?

Principalmente vamos a tener en cuenta la precisión para medir la efectividad del modelo y observar cómo cambia en cada clase.





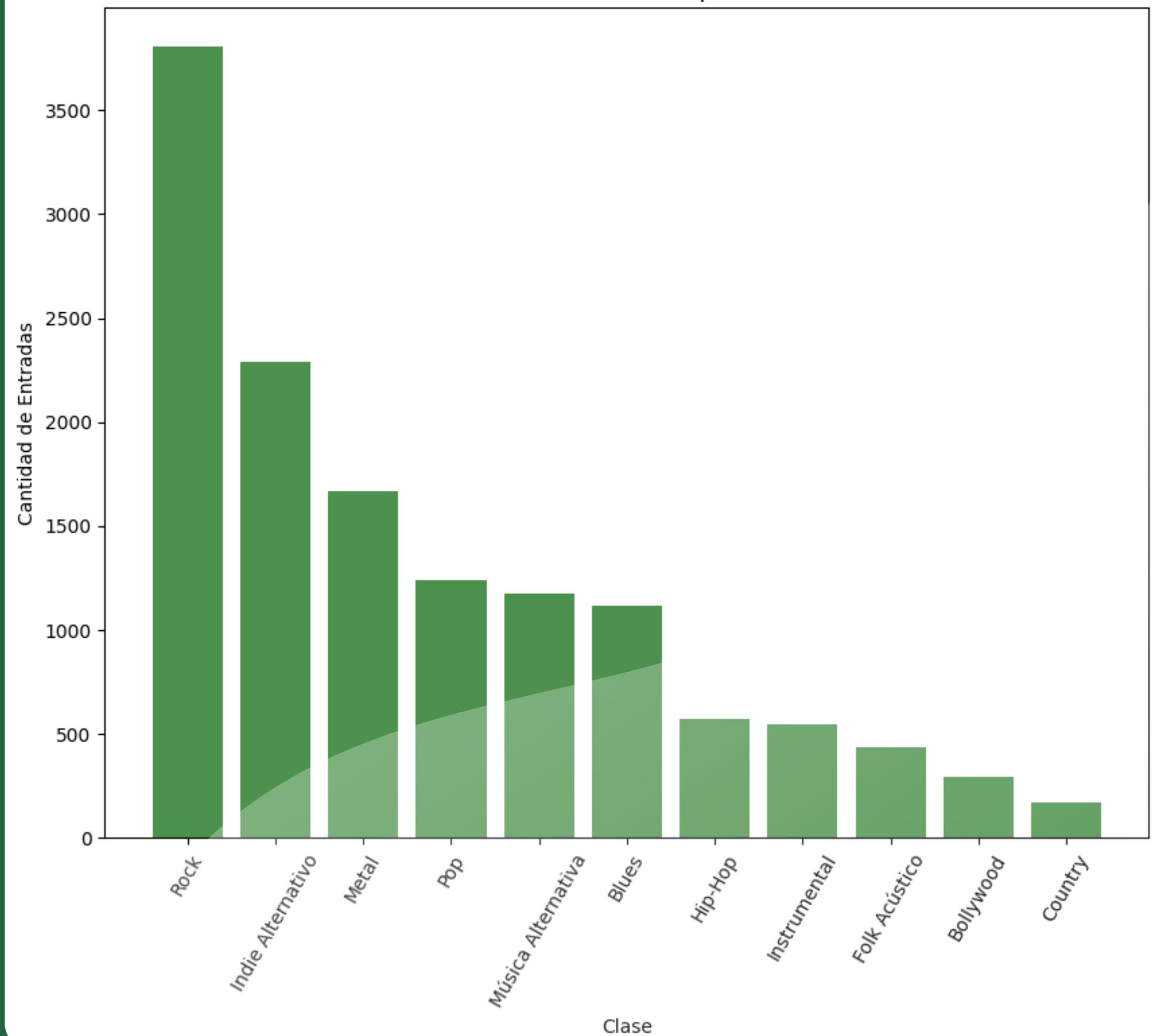
# DESBALANCE

Gran desequilibrio de clases

**Clase Rock**  
3876 entradas  
**MAX**

**Clase Country**  
180 entradas  
**MIN**

Cantidad de Entradas por Clase



# CLASS WEIGHT

Probamos con el parametro **class\_weight** de **RandomForest**:



Modifica la importancia de cada clase a la hora de predecir dependiendo de la cantidad de entradas que tenga.

**'balanced'**

Los pesos se asignan automaticamente

**No produjo efecto en el modelo original**

Train: 0.4126052421137961  
Test: 0.4266032064128257

**Pesos asignados**

Asiganmos los pesos manualmente

# SMOTE

## Over-Sampling con la tecnica SMOTE



Crea ejemplos sintéticos para aumentar la cantidad de entradas de las clases desbalanceadas

Modificamos el dataset

```
smote = SMOTE(sampling_strategy='auto', random_state=42)
```

Cada clase pasa a tener **2.663** entradas

Con los nuevos datos, entrenamos un modelo RandomForest

```
RFC.fit(X_resampled, y_resampled)
```



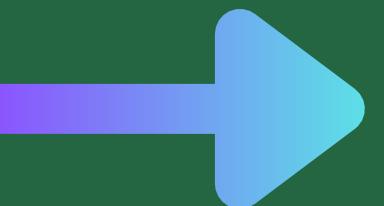


# DF SEPARADO DE ROCK

Los modelos tienden a predecir la mayoría de las clases como Rock debido a la cantidad de entradas que esta tiene respecto a las otras.

**¿Qué pasa si solo observamos Rock?**

Creamos un nuevo modelo con **RandomForest** que solo predice si una entrada pertenece a la clase Rock o no.



## TEST

		Matriz de Confusión	
		Rock	No Rock
Clases reales	Rock	2530	321
	No Rock	892	249
		Rock	No Rock
		Predicciones	

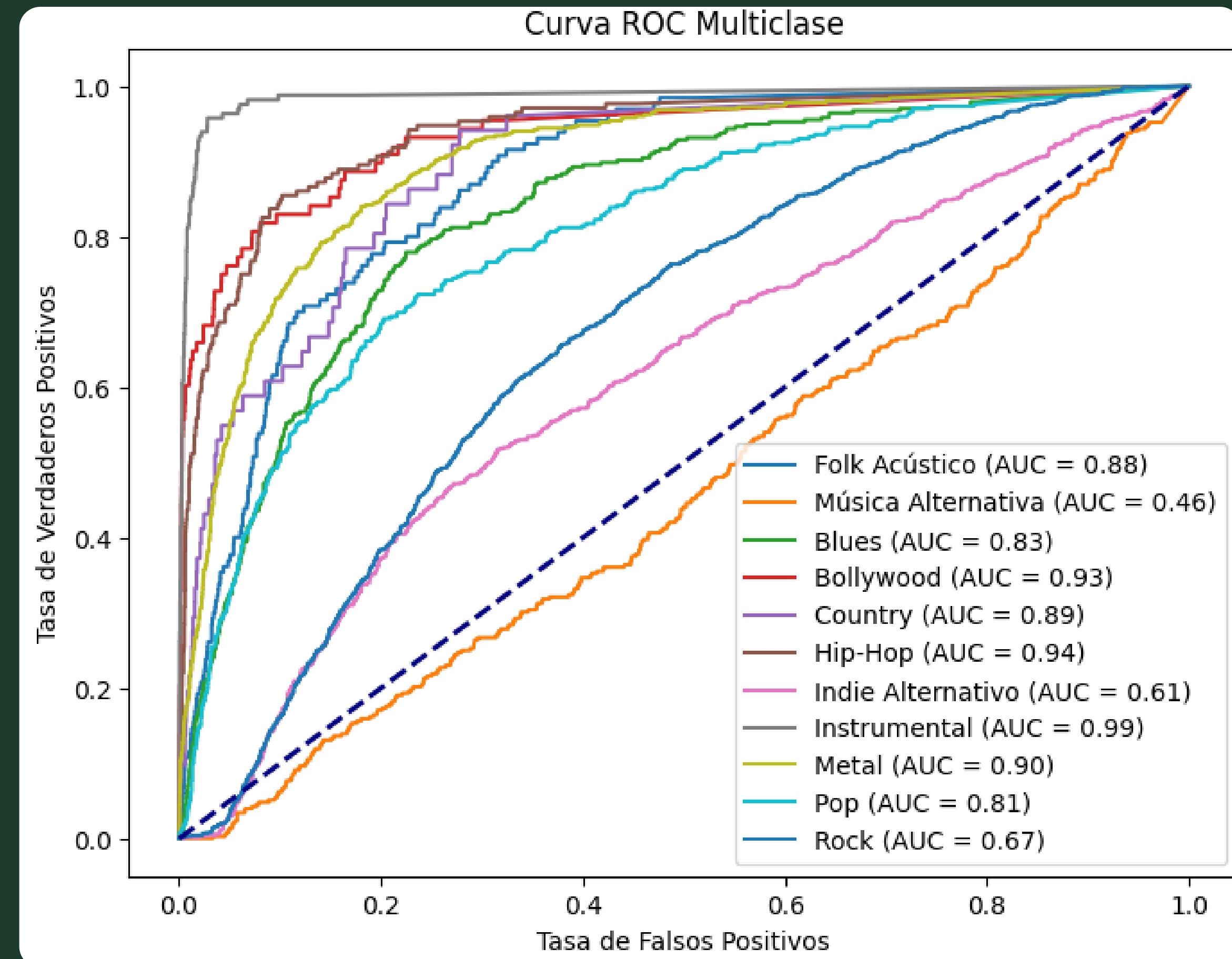
**Accuracy:** 0.69



# CURVA ROC

Vemos las clases que mejor predice el modelo y cuales tienen curvas similares

Curva ROC por cada clase con el modelo RandomForest



# CONCLUSIÓN

- Se logró una predicción decente para la cantidad de clases que había.
- Hay géneros musicales que son más variantes que otros, y por lo tanto menos clasificables.
- Faltan variables o datos para ayudar al modelo a separar las clases cuyas diferencias no son tan evidentes en el dataset actual .

# ELEMENTOS A MEJORAR/AGREGAR

## **Variables musicales nuevas.**

- Instrumentos utilizados
- Temática de la canción
- Fecha de lanzamiento
- Discográfica / Productor
- Disco al que pertenecen

Un conjunto de datos más balanceado con una cantidad de canciones similares por género.

Un dataset más amplio que tenga la intención de incluir todos los géneros musicales, en lugar del actual que está basado únicamente en popularidad.

