

NOV. '24

# ECOS DEL PASADO

ESTRATEGIAS DE CLASIFICACIÓN MUSICAL PARA  
AUMENTAR EL ENGAGEMENT DE USUARIOS DE SPOTIFY

# CONTENIDO

1. Punto de partida

---

2. Entendiendo los datos

---

3. Metodología y modelado

---

4. Modelo elegido

---

5. Un modelo en revisión

---

6. Reflexiones y líneas de mejora

---

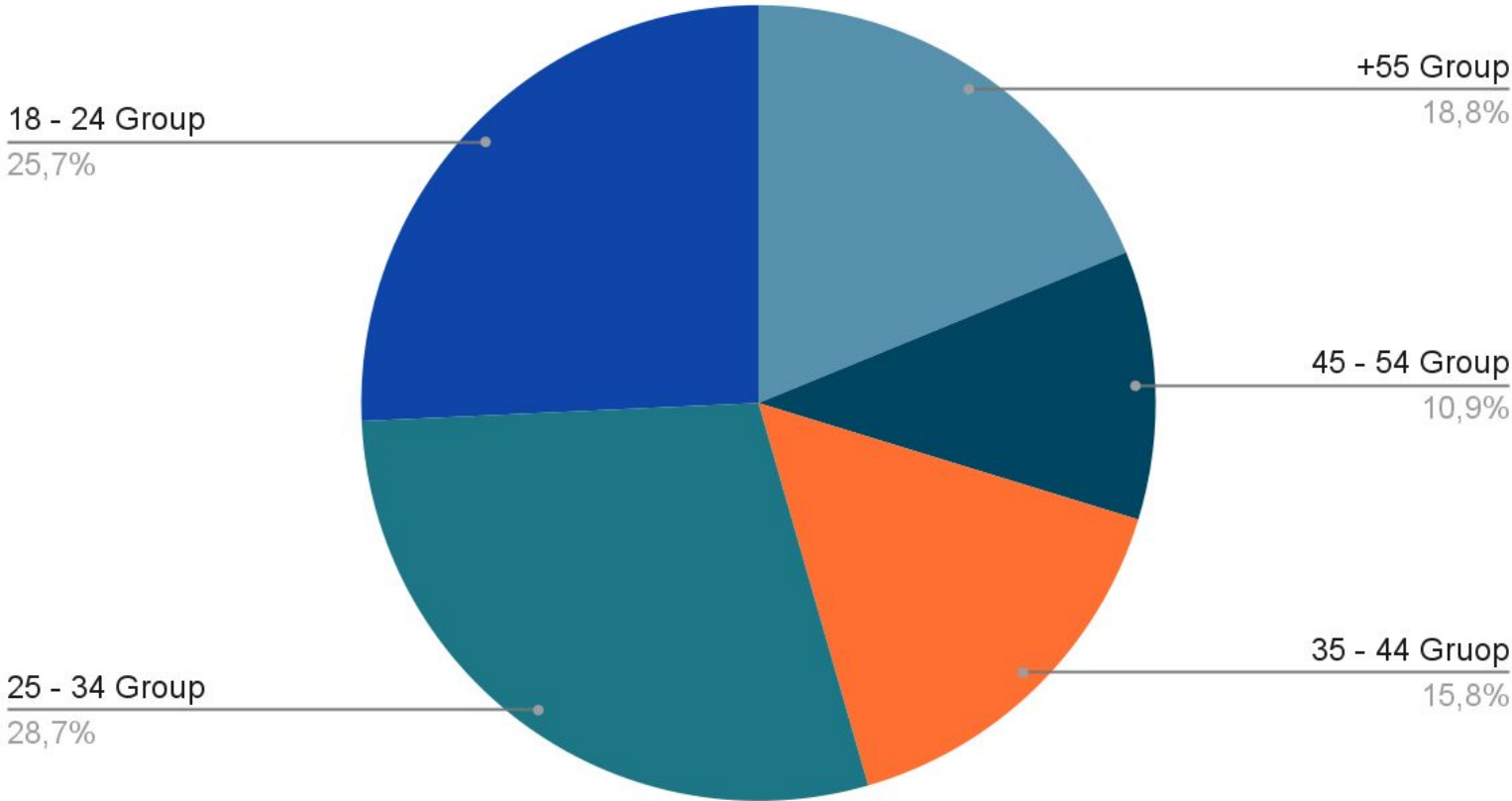
A close-up, low-angle shot of a silver and black dynamic microphone on a stage. The microphone is positioned diagonally from the bottom left towards the top right. The background is dark with several out-of-focus, warm-toned bokeh lights in shades of yellow and orange. The overall mood is dramatic and focused on the microphone as the central subject.

**PUNTO DE PARTIDA**



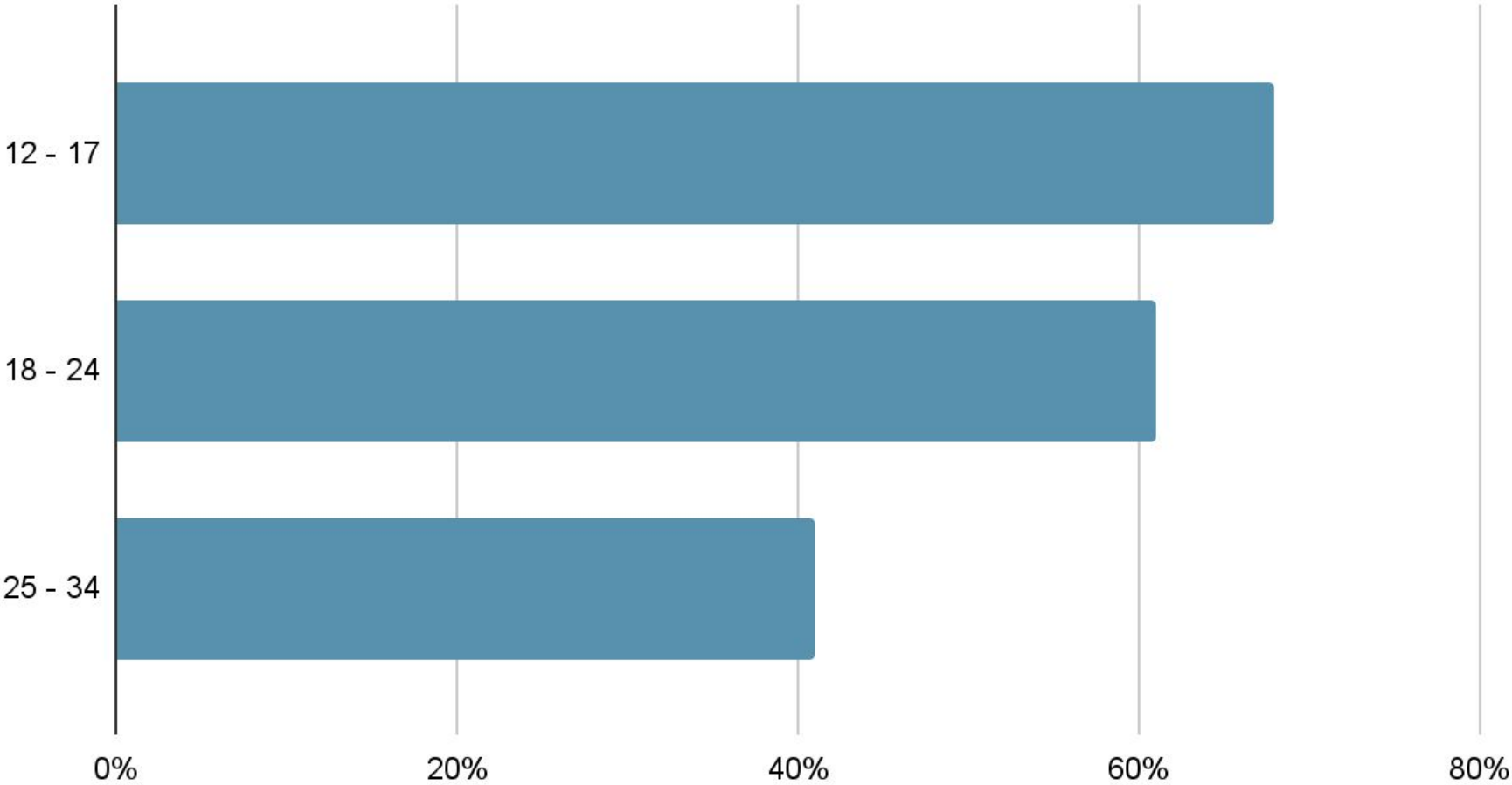
# Un problema a resolver

Age distribution of Spotify users. Global 2023



Referencia: AFFMaven

Porcentaje de usuarios de Spotify en España 2024



Fuente: Statista

**Clasificar canciones modernas según la  
esencia de épocas pasadas para facilitar la  
creación de playlists personalizadas y  
mejorar el engagement de los usuarios de  
edades más avanzadas.**



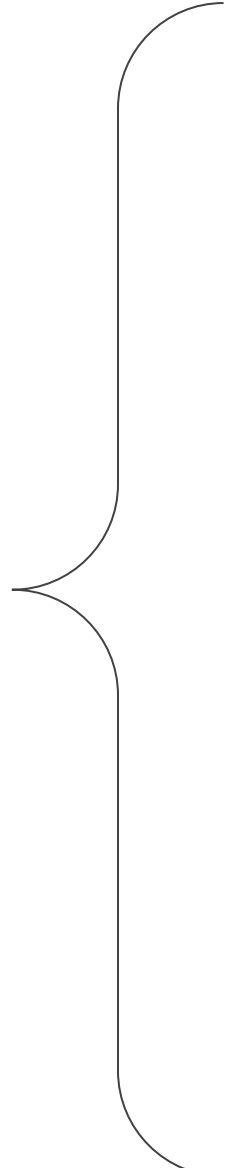
# Dos datasets

## Spotify Songs Dataset

- Canciones entre 1921 y 2020.
- Más de 160K registros.
- Fuente: Kaggle.

## API Spotify Dataset

- Éxitos globales desde 2020 hasta la actualidad.
- 296 registros.
- Fuente: extracción propia a partir de la API de Spotify.

- 
- id
  - name
  - artists
  - duration\_ms
  - release\_date
  - year
  - popularity
  - explicit
  - acousticness
  - danceability
  - energy
  - instrumentalness
  - liveness
  - loudness
  - speechiness
  - tempo
  - valence
  - mode
  - key



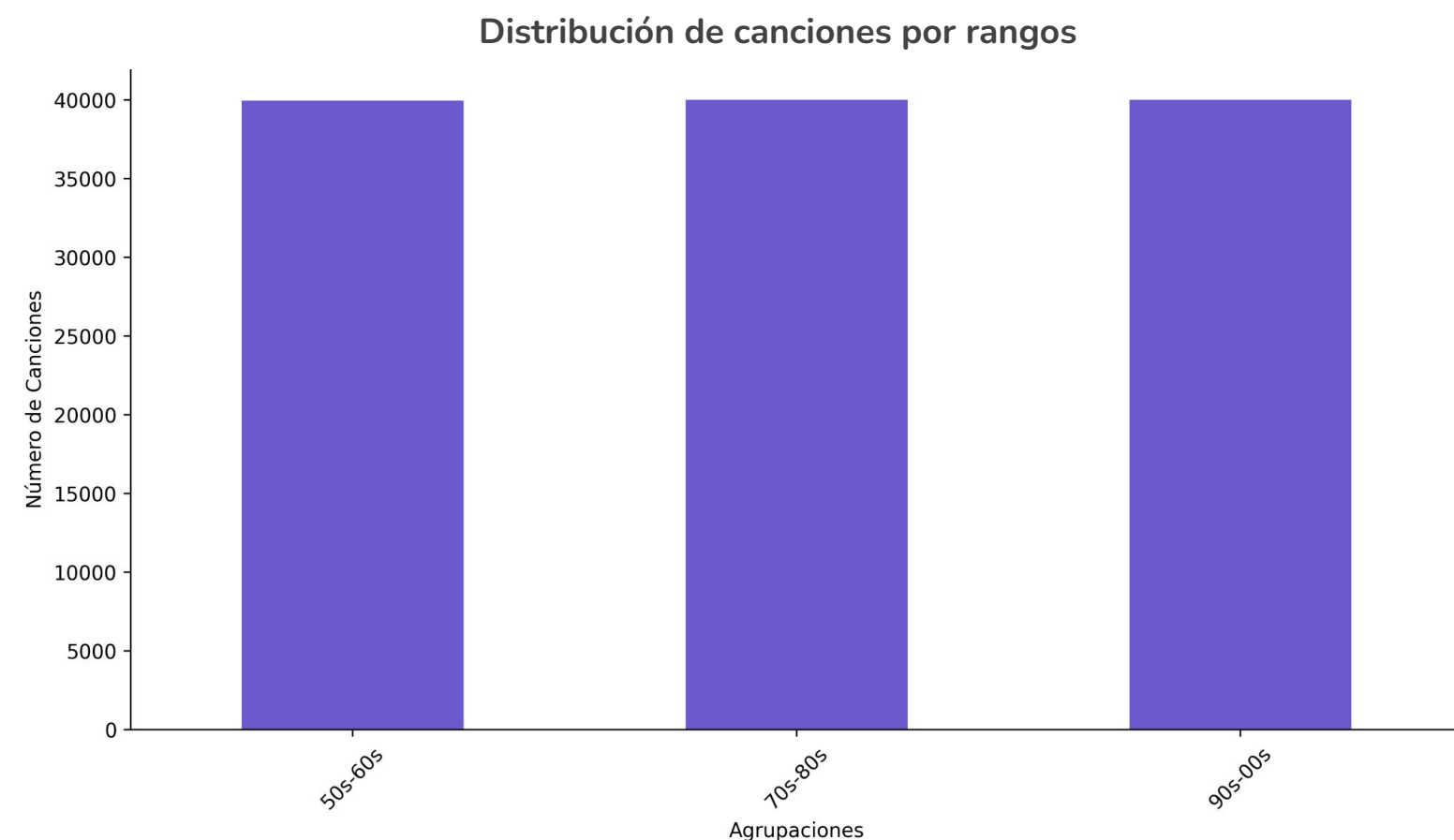
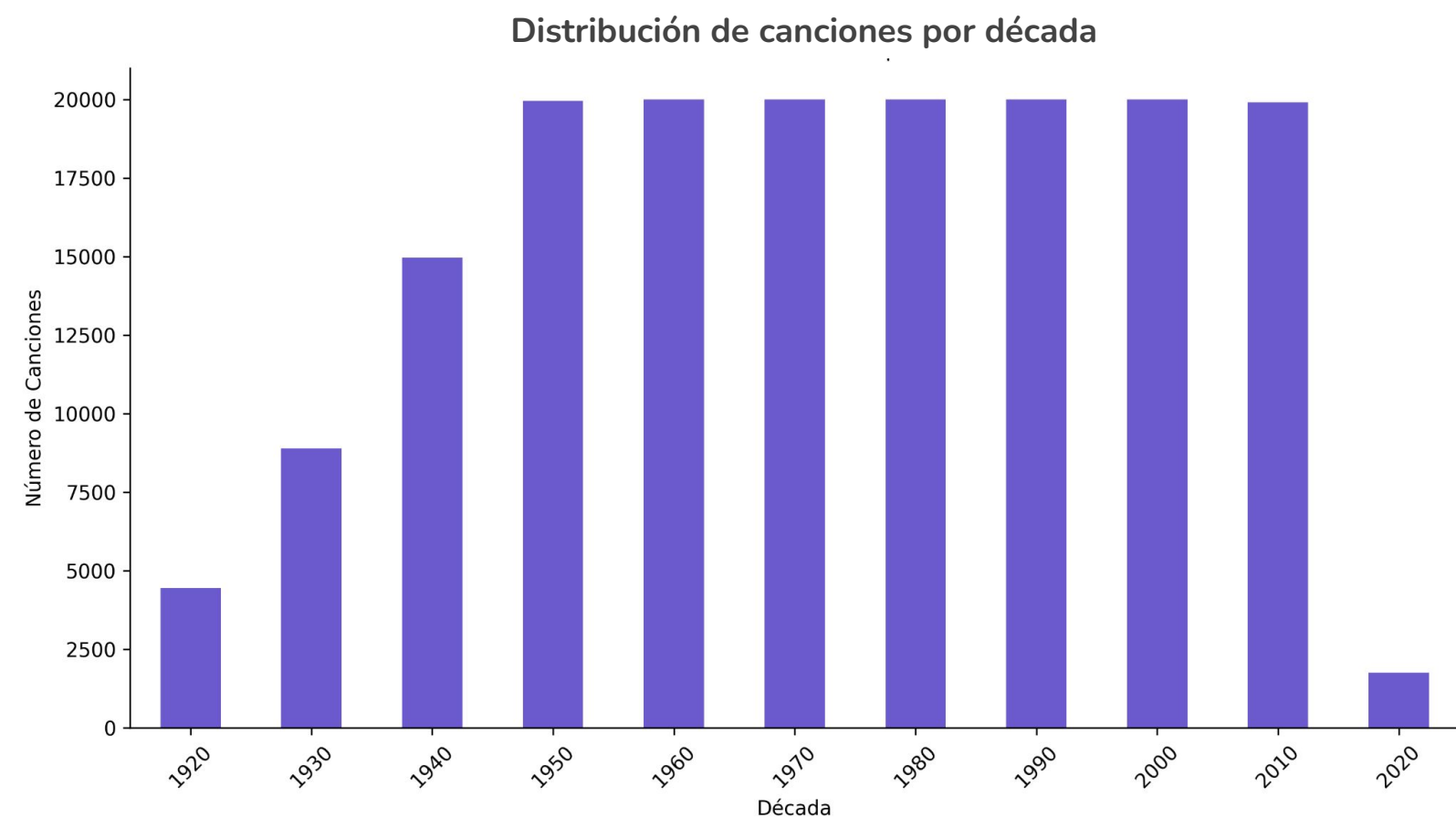


# ENTENDIENDO LOS DATOS



# Trabajando la variable target

Las décadas 50s-60s, 70s-80s, y 90s-00s representan períodos musicales ricos y variados, que abarcan la evolución de géneros y estilos que definieron la música moderna.

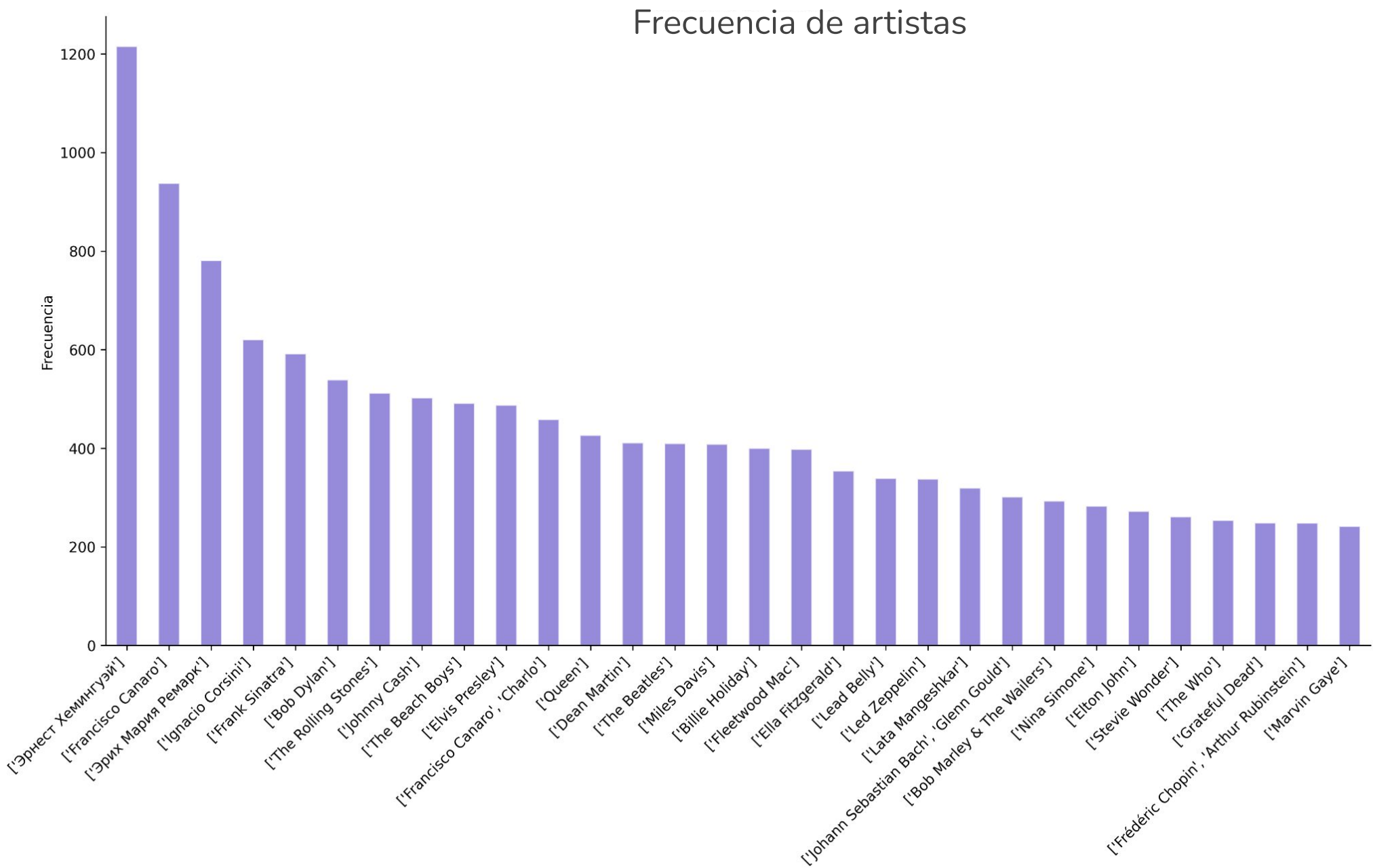






# Prevención del sesgo

La frecuencia de un artista no es un atributo inherente a la música, sino al diseño del dataset. Esto introducía una correlación artificial y llevaba a que el modelo "aprendiera" patrones no generalizables fuera del dataset.





# Nuevas variables

Se profundizó en explorar relaciones que pudieran pasar desapercibidas a los modelos mediante la creación de nuevas variables a partir de las existentes.

**"Energíaailable positiva":** La combinación de energy, danceability, y valence podría representar canciones energéticas,ailables y de tono positivo, características comunes en ciertas décadas.

**"Intensidad acústica":** Multiplicar acousticness con loudness podría crear una métrica que mida la intensidad de las canciones acústicas, lo que podría ser útil para identificar canciones de épocas con un sonido más orgánico.

**Ratio de popularidad con energía:** La relación entre popularity y energy podría ser útil para medir si una canción energética tiene una mayor o menor popularidad, lo que podría variar entre décadas.

**Diferencia entre valence y energy:** Este valor podría ayudar a diferenciar canciones alegres y energéticas de canciones tranquilas pero alegres para explorar si ciertas décadas eran más propensas a propuestas de este tipo.

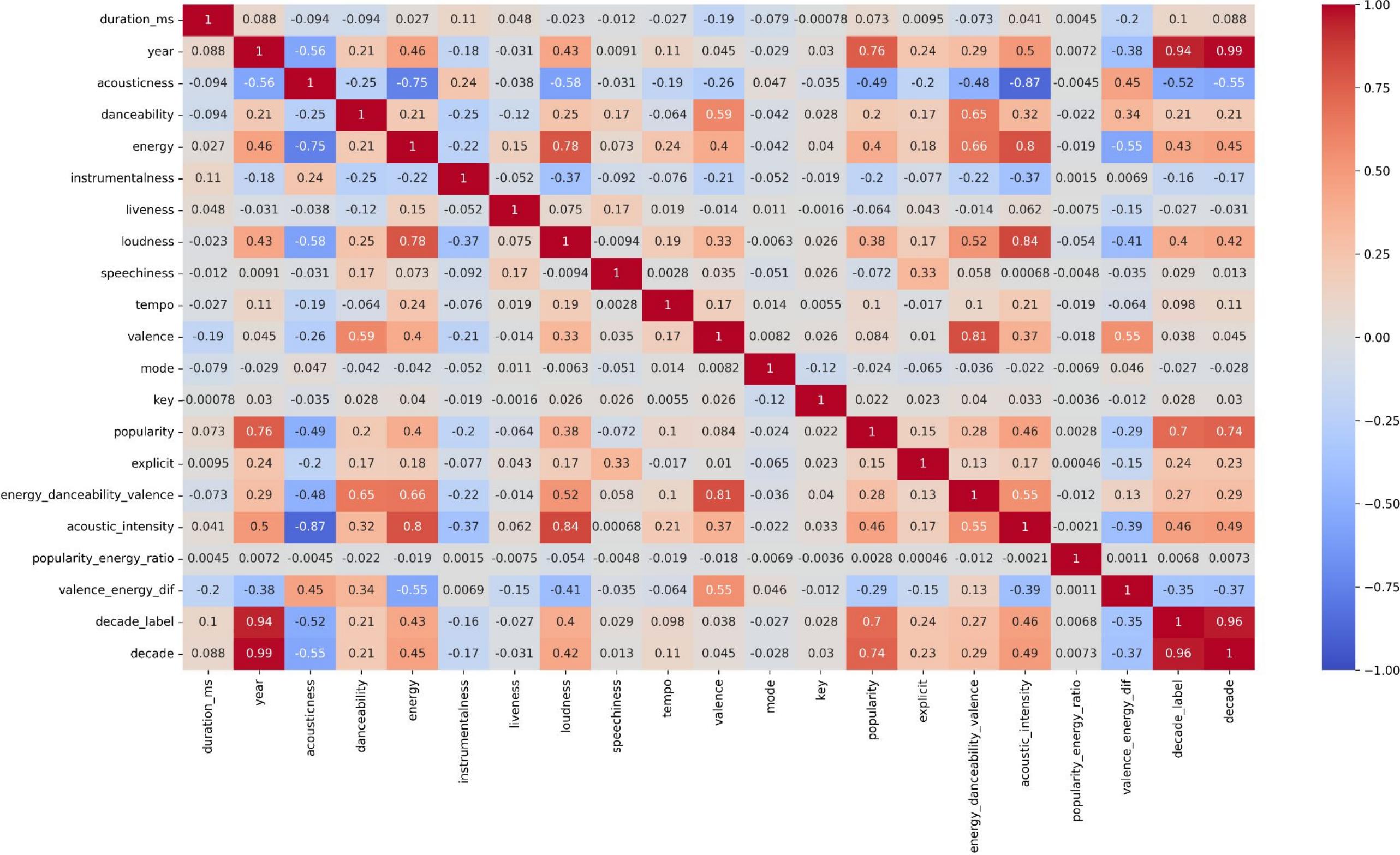


**Nuevas  
variables**

**Variables  
originales**



# Correlaciones



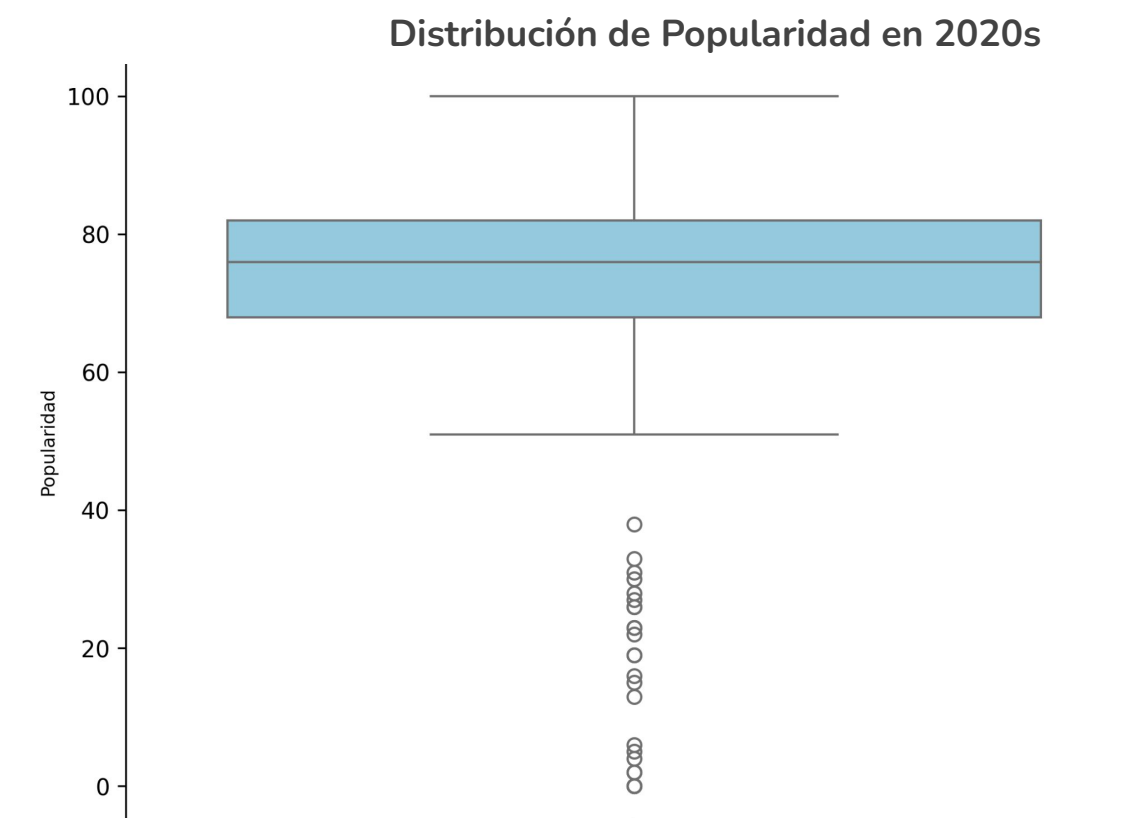
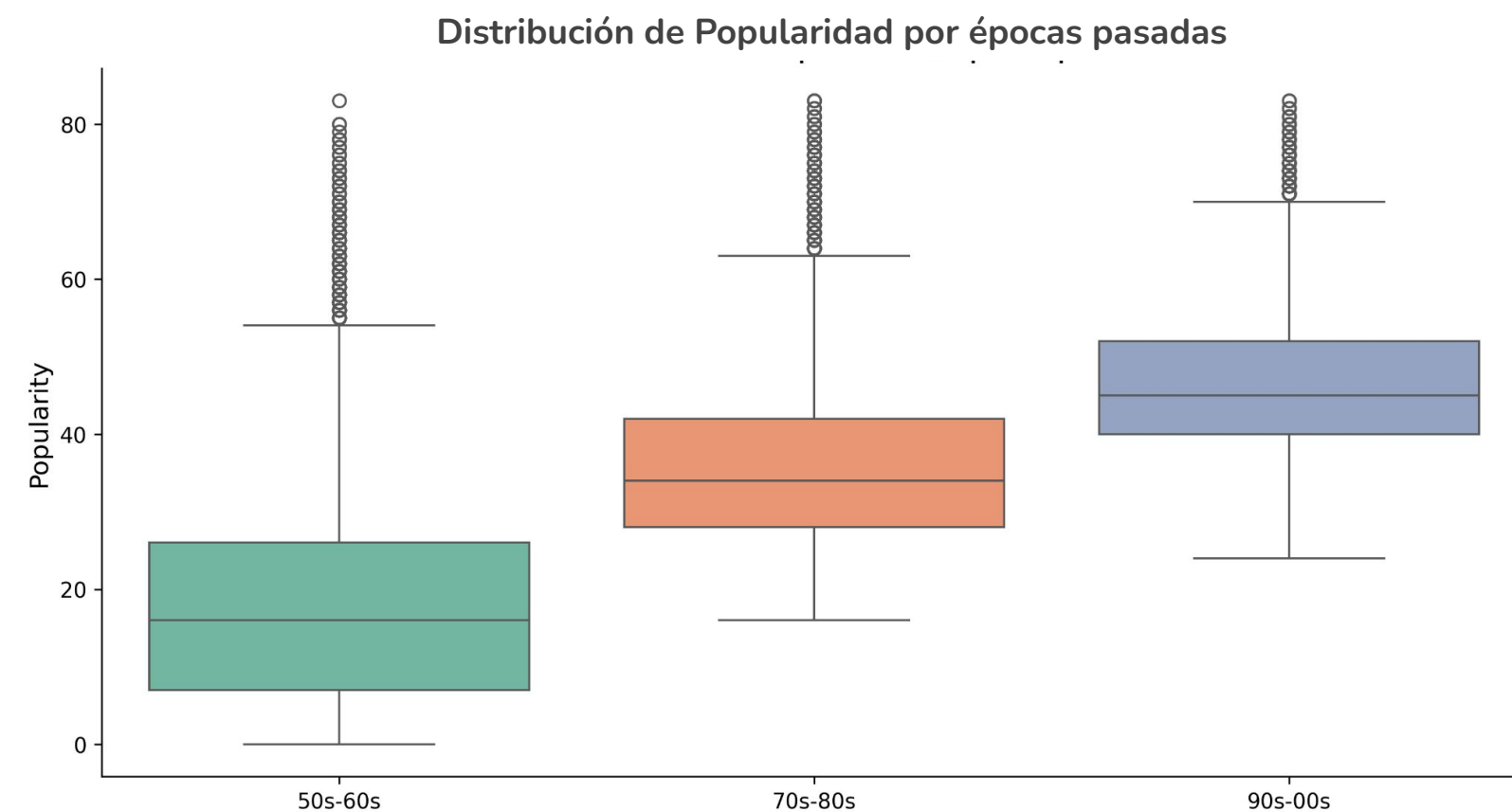






# Normalizando la popularidad

El comportamiento de la variable popularity se demostró crítica, introduciendo un sesgo significativo en el modelo, ya que no representa un atributo inherente de las canciones, sino su contexto actual. Se normalizó la popularidad en función de cada época.





# METODOLOGÍA Y MODELADO

## DecisionTree

```
{  
- N° variables: 11,  
- Feature importance: Sí,  
- Tipo de variables:  
  originales y nuevas,  
- Mejores parámetros:  
  - 'max_depth': None,  
  - 'min_samples_leaf': 4,  
  - 'min_samples_split': 10,  
}
```

## RandomForest

```
{  
- N° variables: 6,  
- Feature importance: Sí,  
- Tipo de variables:  
  originales,  
- Hiperparametrización: No  
}
```

## SVM

```
{  
- Sample size: 20.000,  
- N° variables: 14,  
- SelecKBest: Sí,  
- Tipo de variables:  
  originales y nuevas.  
- Escalado: RobustScaler,  
- Hiperparametrización:  
  - 'C': 1,  
  - 'gamma': 0.1,  
  - 'kernel': 'rbf'  
}
```



## KNN

```
{  
- N° variables: 9,  
- Feature importance: Sí  
  (XGB),  
- Tipo de variables:  
  originales,  
- Escalado: StandardScaler,  
- Pre-procesado: PCA (8),  
- Hiperparametrización: No  
}
```

## Red Neuronal

```
{  
- Keras: Sequential,  
- N° variables: 21,  
- Capas densas: 3 (128, 64,  
  16) (ReLU)  
- Dropout: 30%  
- Optimizador: Adam,  
- Función de pérdida:  
  categorical_crossentropy  
- Tipo de variables:  
  originales y nuevas.  
}
```

## Pipeline

```
{  
- N° variables: 23.  
- Tipo de variables:  
  originales y nuevas.  
-  
  SVM:  
  'svm__C': 10,  
  'svm__gamma': 'scale',  
  'svm__kernel': 'rbf'  
  
  KNN:  
  'knn__n_neighbors': 10,  
  'knn__p': 1,  
  'knn__weights': 'distance'  
}
```



**MODELO  
ELEGIDO**



# XGBoost

```
{'learning_rate': 0.1, 'max_depth': 5,  
'n_estimators': 100, 'reg_alpha': 0,  
'reg_lambda': 1}
```

## Rendimiento en datos no vistos

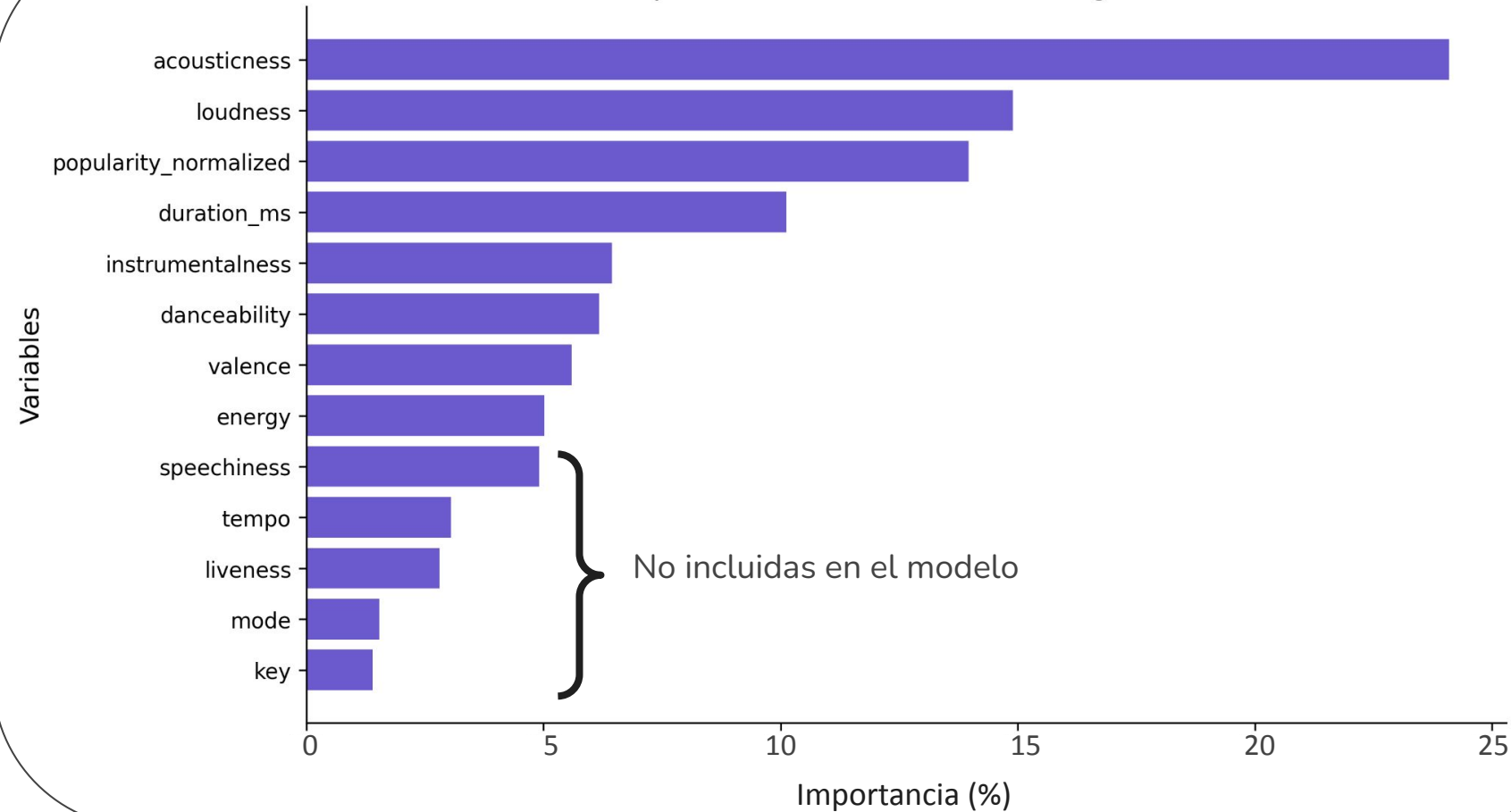
**0.885**  
Accuracy

**0.887**  
Precision

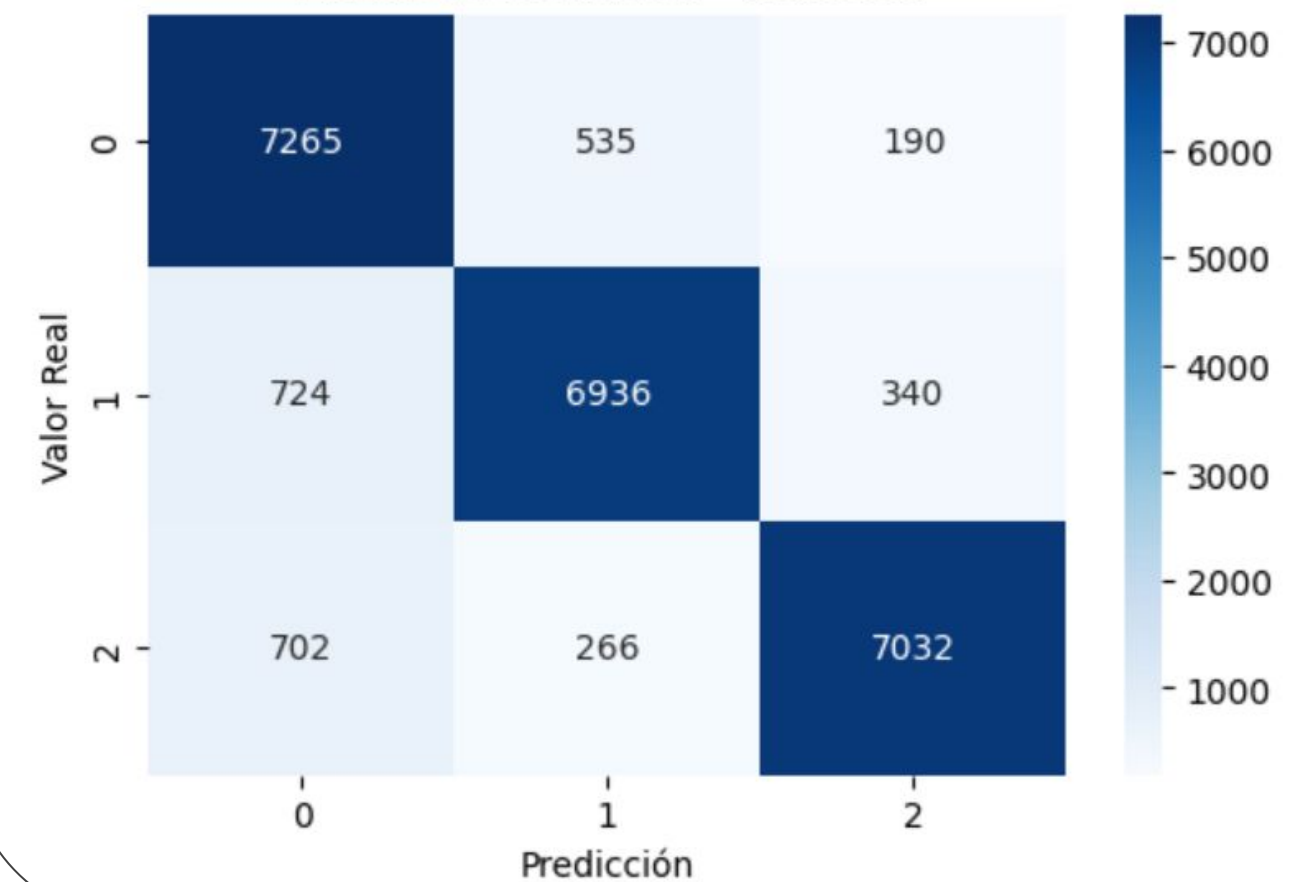
**0.885**  
Recall

**0.885**  
F1-Score

Importancia de las variables originales



Matriz de Confusión - XGBoost



## Reporte de clasificación en datos no vistos

**50s-60s**

**0.84**  
Precision

**0.91**  
Recall

**0.87**  
F1-Score

**70s-80s**

**0.90**  
Precision

**0.87**  
Recall

**0.88**  
F1-Score


**90s-00s**

**0.93**  
Precision

**0.88**  
Recall

**0.90**  
F1-Score





**UN MODELO  
EN REVISIÓN**



# XGBoost

## Rendimiento en datos no vistos

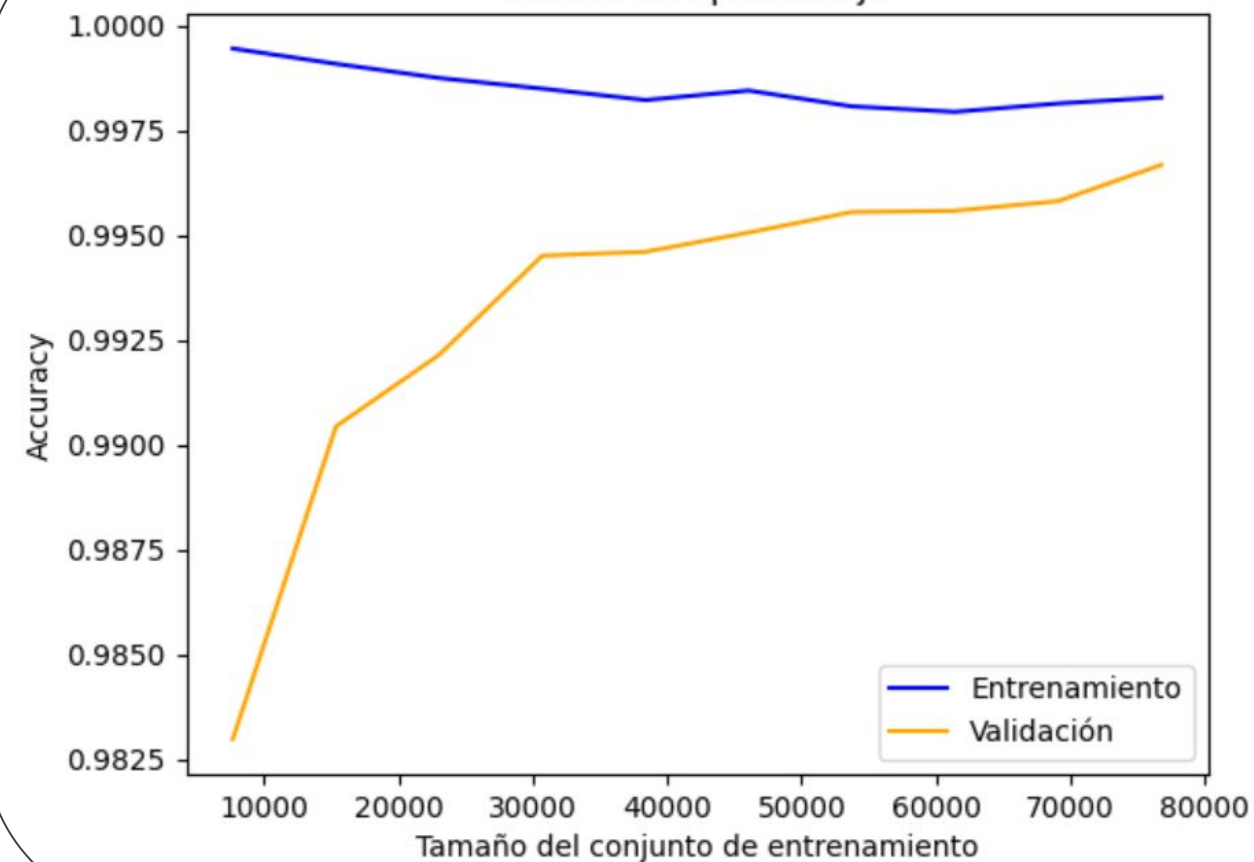
**0.997**  
Accuracy

**0.997**  
Precision

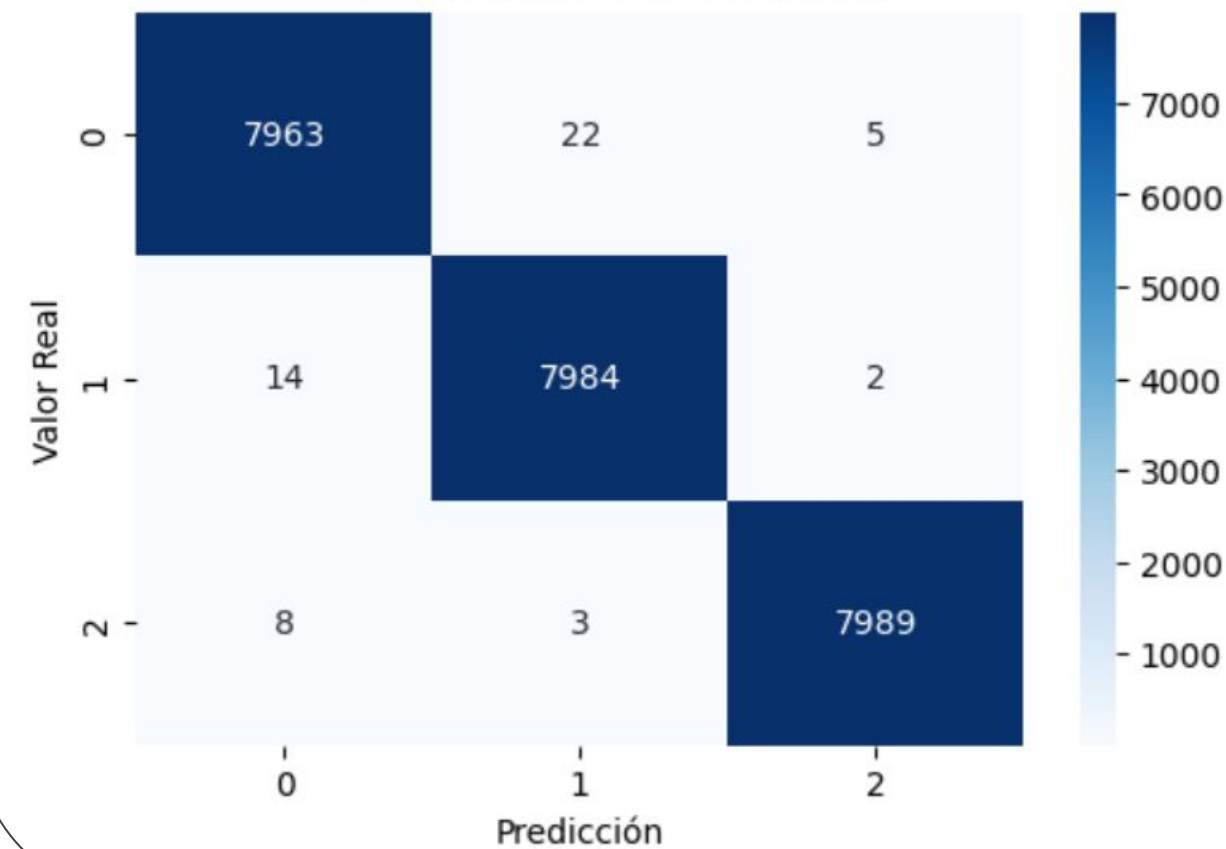
**0.997**  
Recall

**0.997**  
F1-Score

Curvas de aprendizaje



Matriz de Confusión - XGBoost



## Cross-Validation (CV=5)

- 0.99124635
- 0.99270529
- 0.99291371
- 0.99249687
- 0.99458108

**Media de scores:**  
0.99278

## Reporte de clasificación en datos no vistos

### 50s-60s

**1.00**  
Precision

**1.00**  
Recall

**1.00**  
F1-Score

### 70s-80s

**1.00**  
Precision

**1.00**  
Recall

**1.00**  
F1-Score

### 90s-00s

**1.00**  
Precision

**1.00**  
Recall

**1.00**  
F1-Score

# XGBoost

	name	artists	year	predicted_decade
0	Multiply (feat. Nate Dogg)	['Xzibit', 'Nate Dogg']	2002	2
1	Un Millón De Lágrimas	['Tropical Panamá']	1992	2
2	Rhayader Goes To Town	['Camel']	1975	1
3	We Suck Young Blood	['Radiohead']	2003	2
4	Cariño Santo - Version 1980	['Los Baron De Apodaca']	1980	1
5	Color Me True	['Sly & The Family Stone']	1968	0
6	Turn! Turn! Turn!	['The Byrds']	1990	2
7	She Sells	['Roxy Music']	1975	1
8	Adventures In Paradise	['Manny Lopez']	1961	0
9	People Make The World Go Round	['Milt Jackson']	1973	1
10	Walking Away	['Jonny Lang']	1998	2
11	Circles	['Joe Satriani']	1987	1
12	Popurri	['Aniceto Molina']	2005	2
13	Kako Y Palmieri	['Alegre All Stars']	1961	0
14	La Gorra	['Tropical Del Bravo']	1999	2
15	Debussy: Mazurka, L. 75, L. 67	['Claude Debussy', 'Walter Giesecking']	1953	0
16	Can't Fake the Feeling - Radio Edit	['Geraldine Hunt']	1980	1
17	The Weight - Live At The Fillmore East/1970	['Joe Cocker']	1970	1
18	Roots and Culture	['Mikey Dread']	1998	2



# REFLEXIONES Y LÍNEAS DE MEJORA



# Reflexiones y líneas de mejora

- Dataset de extracción propia intentando reducir los sesgos.
- Nuevas líneas de investigación y personalización. Por ejemplo:
  - Clasificación por épocas pero en diferentes territorios.
  - Clasificación de canciones/artistas actuales con la esencia de artistas de otras épocas.
- Refinar modelos actuales.
- Seguir revisando el “modelo sobresaliente”.





**PARA  
TERMINAR**



# ¿Lo probamos?



Streamlit

Predecir



NOV. '24

# ¡GRACIAS!