

ANÁLISIS DE CANCIONES EN SPOTIFY

TRABAJO PRÁCTICO FINAL - ANÁLISIS DE DATOS

Luciano Adassus
Ignacio Tomas de Pedro Mermier
Jonathan Cagua Ordoñez
Agustina Quiros

OBJETIVO DEL PROYECTO

Estimar la probabilidad de que una nueva canción sea del agrado de un usuario, basado en las características de su playlist actual.

LOS DATOS

El dataset utilizado en este análisis fue obtenido de una colección de canciones en Spotify.

[DATASET EN GOOGLE DRIVE](#)

EL DATASET

TAMAÑO

El dataset cuenta con 750 registros y 14 columnas.

TIPOS DE DATOS

La mayoría de las variables son de tipo float64 (9 variables). El resto de las variables son de tipo int64 (5 variables).

CLASIFICACIÓN DE VARIABLES

VARIABLES NUMÉRICAS

Continuas: acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence. Discretas: duration.

VARIABLES CATEGÓRICAS

Estas variables agrupan los datos en categorías y se dividen en **nominales**, **ordinales** y **binarias**. Nominales: key, mode. Ordinales: time_signature. Binarias: label.

spotify_df.info()			
[152]	Column	Non-Null Count	Dtype
0	acousticness	750 non-null	float64
1	danceability	750 non-null	float64
2	duration	750 non-null	int64
3	energy	750 non-null	float64
4	instrumentalness	750 non-null	float64
5	key	750 non-null	int64
6	liveness	750 non-null	float64
7	loudness	750 non-null	float64
8	mode	750 non-null	int64
9	speechiness	750 non-null	float64
10	tempo	750 non-null	float64
11	time_signature	750 non-null	int64
12	valence	750 non-null	float64
13	label	750 non-null	int64

dtypes: float64(9), int64(5)
memory usage: 82.2 KB

Variable	Tipo	Descripción
acousticness	Continua	Atributo que mide qué tan acústica es una canción.
danceability	Continua	Mide lo apta que es una canción para bailar.
duration	Discreta	Duración de una canción en milisegundos.
energy	Continua	Mide la intensidad y actividad percibida de la canción.
instrumentalness	Continua	Indica la probabilidad de que una pista no contenga voces.
key	Discreta	Representa la tonalidad musical de una canción (en semitonos).
liveness	Continua	Detecta la presencia de una audiencia en la grabación.
loudness	Continua	Mide el volumen medio de la canción (en decibelios).
mode	Binaria	Indica si una pista está en modo mayor (1) o menor (0).
speechiness	Continua	Mide la cantidad de palabras habladas en la pista.
tempo	Continua	Mide el ritmo de la canción en beats por minuto (BPM).
time_signature	Ordinal	Representa la métrica de la canción (por ejemplo, 4/4).
valence	Continua	Mide el carácter musical de una pista en términos de positividad.
label	Binaria	Variable objetivo que se desea predecir.

Anomalías en variables

SE IDENTIFICAN VALORES FUERA DE LOS RANGOS TÍPICOS PARA LAS VARIABLES IMPORTANTES COMO: DURATION, TEMPO, Y LOUDNESS. ESTOS VALORES PUEDEN SER INDICATIVOS DE OUTLIERS QUE PODRÍAN AFECTAR LOS ANÁLISIS POSTERIORES.

```
# Definimos los rangos esperados para algunas variables clave
expected_ranges = {
    'tempo': (50, 200), # BPM típicos
    'duration': (60000, 600000), # Entre 1 y 10 minutos
    'loudness': (-60, 0) # Valores de decibelios típicos
}
```

Anomalías encontradas en tempo:

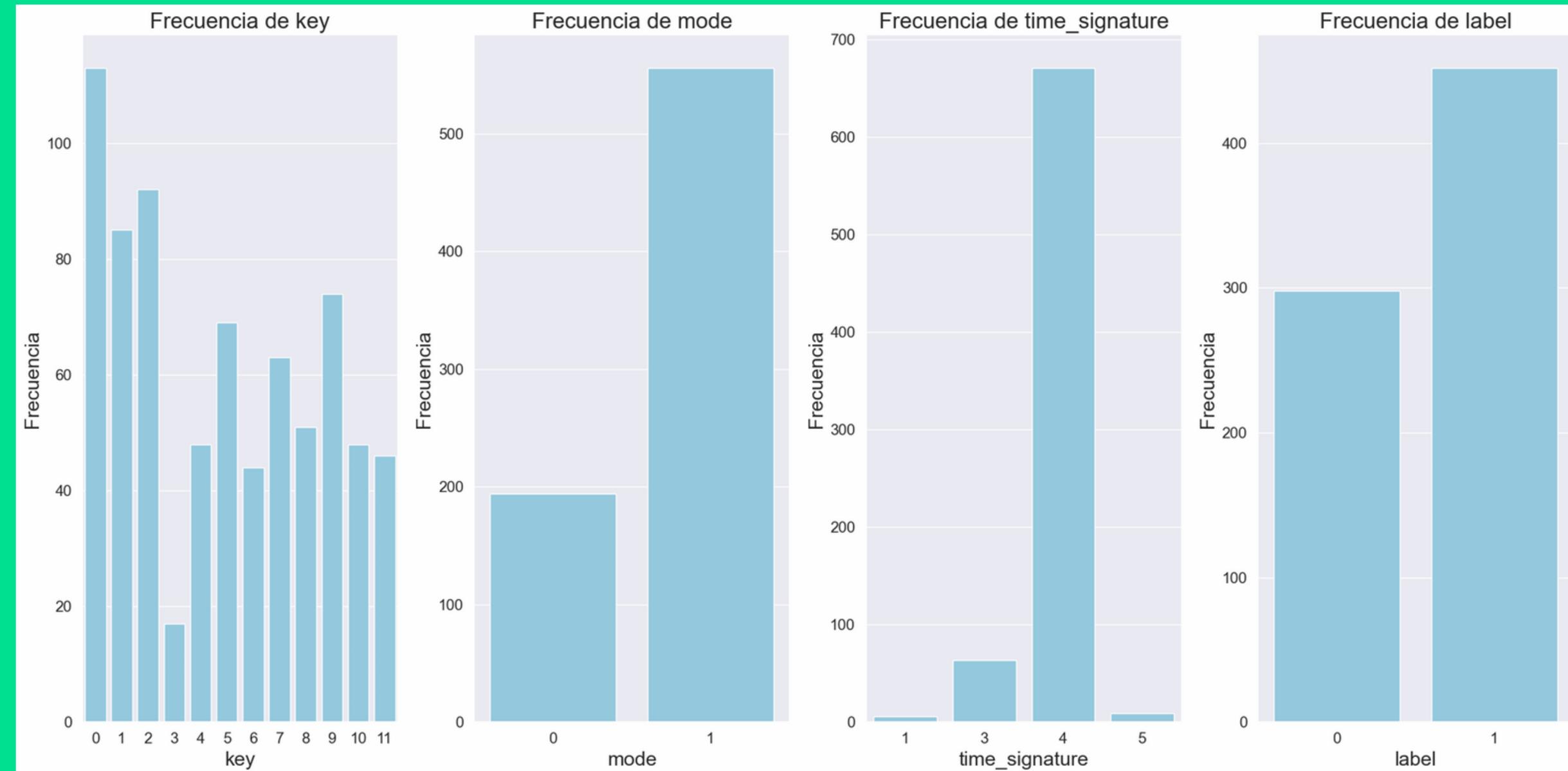
	tempo
195	201.843
350	203.927
377	203.669
649	201.800
654	203.988
743	204.162

Anomalías encontradas en duration:

	duration
215	46107
241	675360
351	55653
449	58671
488	33840
651	56331
700	48093
730	618400

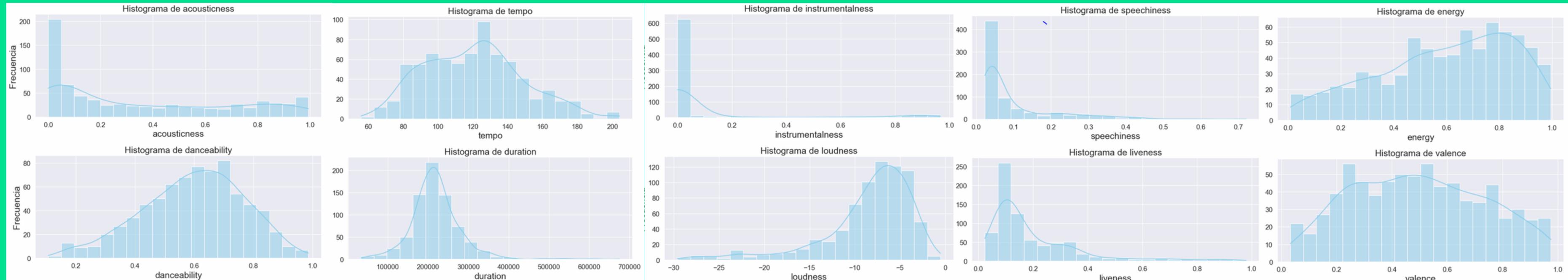
No se encontraron anomalías en loudness.

DISTRIBUCIONES DE VARIABLES



Categóricas

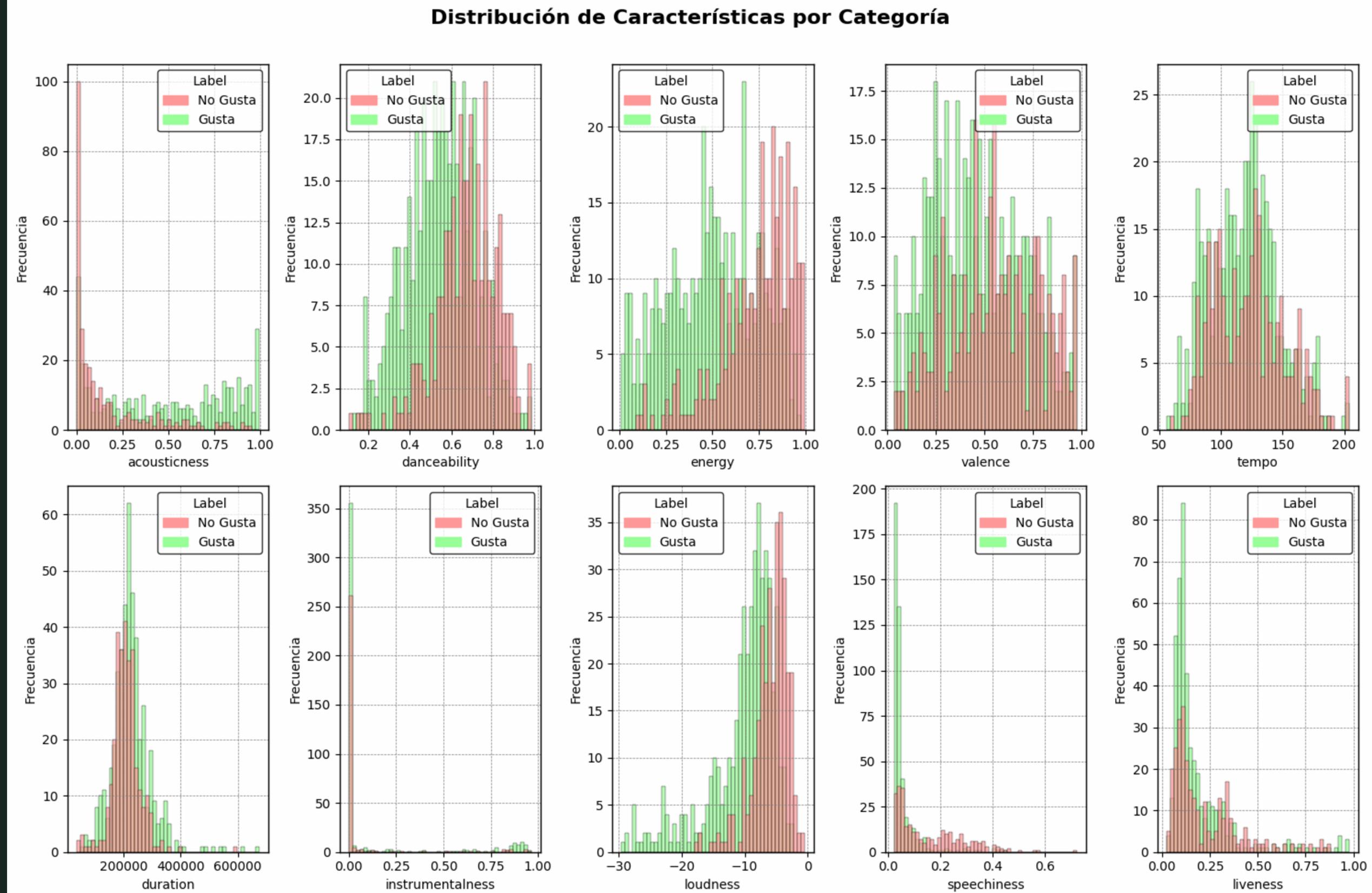
DISTRIBUCIONES DE VARIABLES



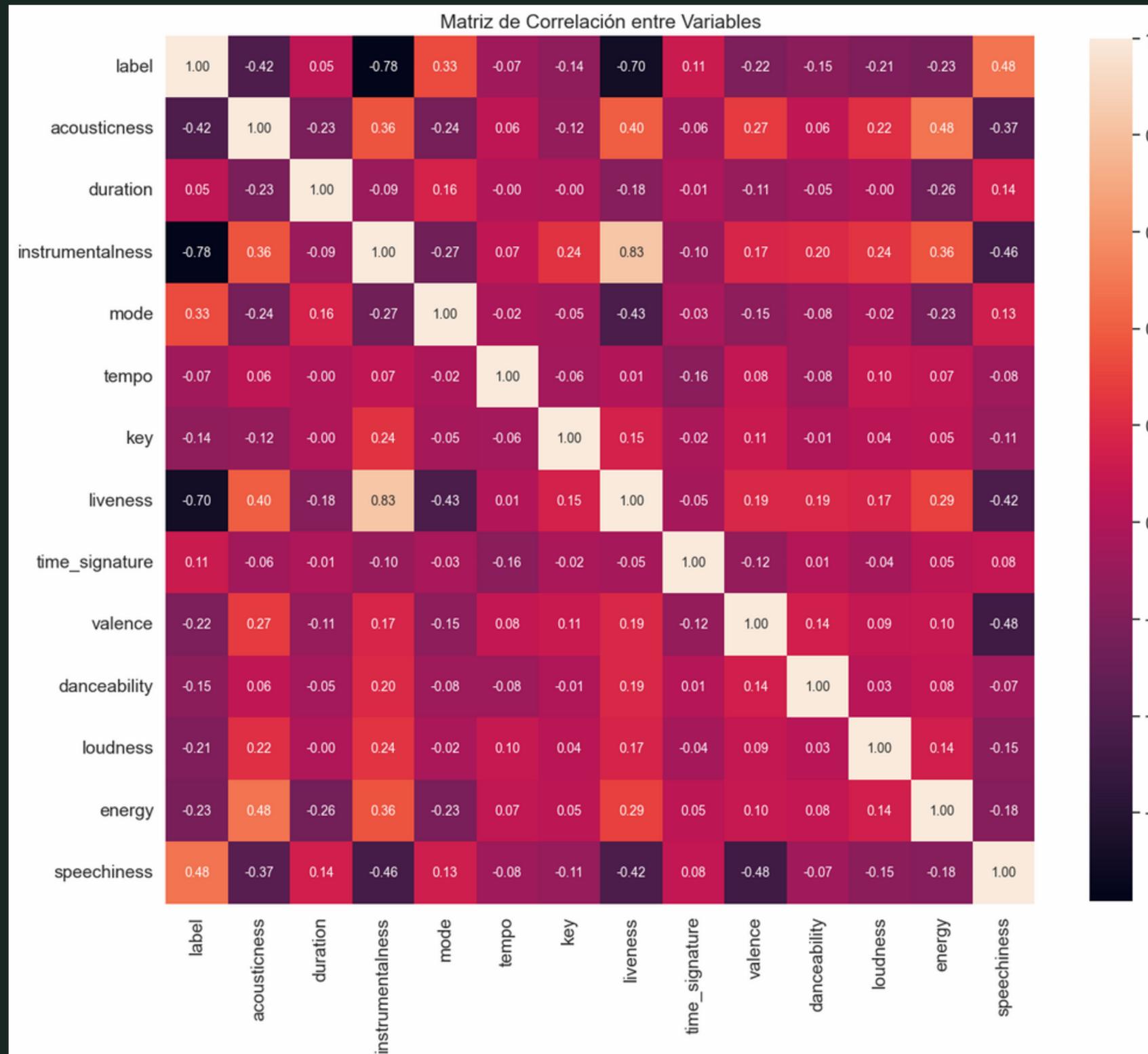
Numéricas

Aplicación de Técnicas de Visualización

Distribución de Características por Categoría



Correlaciones



Definimos un umbral de 0.9 para identificar aquellas variables que comparten una cantidad significativa de información pero que no son completamente redundantes.

FEATURES ELIMINADAS

No se observaron correlaciones significativas mayores a 0.9, por el momento, no desestimamos ningún feature del dataset.

VALORES NULOS Y DUPLICADOS

	Cantidad nulls
acousticness	0
danceability	0
duration	0
energy	0
instrumentalness	0
key	0
liveness	0
loudness	0
mode	0
speechiness	0
tempo	0
time_signature	0
valence	0
label	0

```
len(spotify_df[spotify_df.duplicated()])  
✓ 0.0s  
14
```

Se identificaron 14 registros duplicados los cuales fueron eliminados.

	acousticness	danceability	duration	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence	label	
83	0.046900	0.311	208467	0.3250	0.000000	2	0.1390	-9.042	1	0.0283	65.090		1	0.668	1
151	0.586000	0.565	238933	0.4610	0.000000	0	0.1620	-7.273	1	0.1410	157.894		4	0.199	1
233	0.992000	0.525	228253	0.0553	0.933000	2	0.0934	-22.358	1	0.0633	67.325		4	0.256	1
263	0.025400	0.541	205200	0.8540	0.000125	2	0.6510	-6.196	1	0.1550	86.044		4	0.454	0
297	0.992000	0.525	226293	0.0633	0.905000	9	0.1050	-23.072	1	0.0497	71.855		4	0.297	1
301	0.586000	0.565	238933	0.4610	0.000000	0	0.1620	-7.273	1	0.1410	157.894		4	0.199	1
426	0.182000	0.874	216248	0.7060	0.000000	1	0.3340	-5.132	1	0.2070	89.968		4	0.895	0
450	0.166000	0.708	213440	0.6660	0.000229	2	0.0929	-7.042	1	0.0349	89.019		4	0.834	1
537	0.137000	0.666	211931	0.9480	0.000000	10	0.1920	-2.776	1	0.0638	100.996		4	0.523	0
542	0.166000	0.708	213440	0.6660	0.000229	2	0.0929	-7.042	1	0.0349	89.019		4	0.834	1
547	0.849000	0.390	184667	0.3020	0.000191	0	0.1220	-10.362	1	0.0379	109.394		3	0.232	1
564	0.000986	0.578	188013	0.8250	0.000000	1	0.1760	-6.107	1	0.3220	130.089		4	0.283	0
602	0.005870	0.825	220627	0.8320	0.000789	5	0.1140	-5.853	0	0.0403	122.021		4	0.713	1
662	0.002130	0.733	293543	0.5430	0.000169	1	0.0703	-10.002	1	0.0445	106.019		4	0.118	0

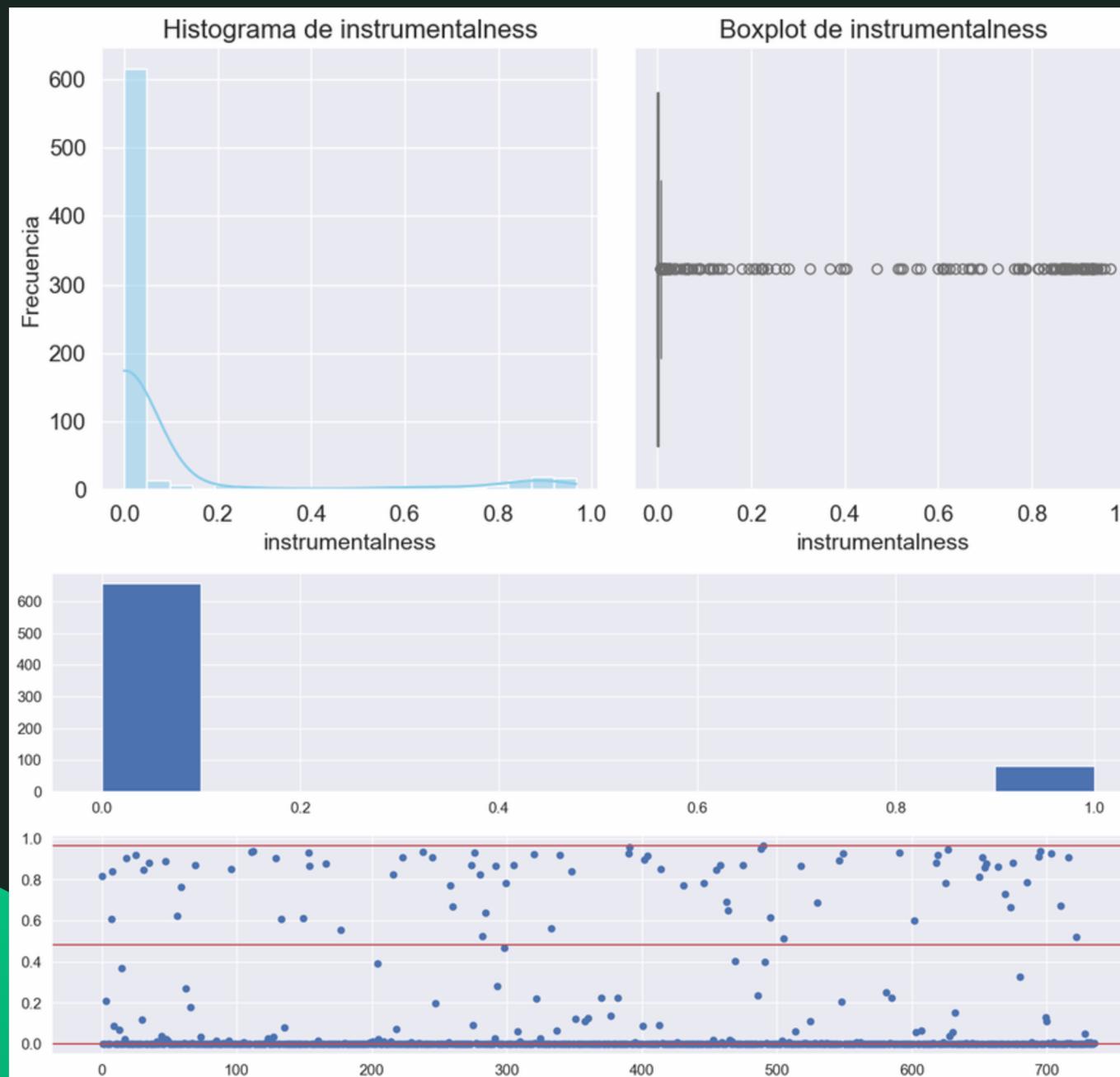
```
spotify_df_filter.shape  
✓ 0.0s  
(736, 14)
```

DATASET LUEGO DE LA LIMPIEZA

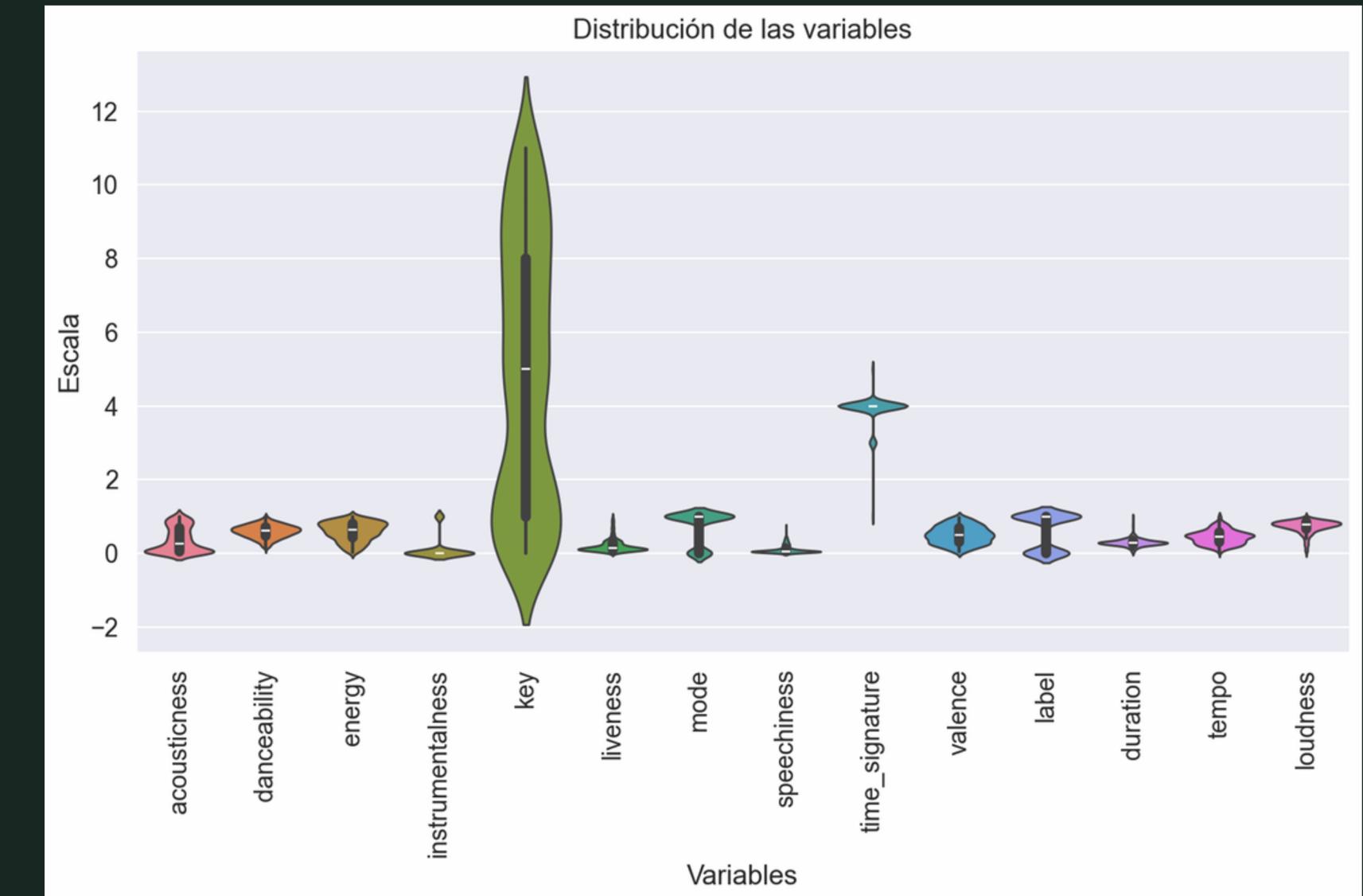


DISCRETIZACIÓN Y ESCALAMIENTO

Discretización uniforme de “instrumentalness” a una variable binaria



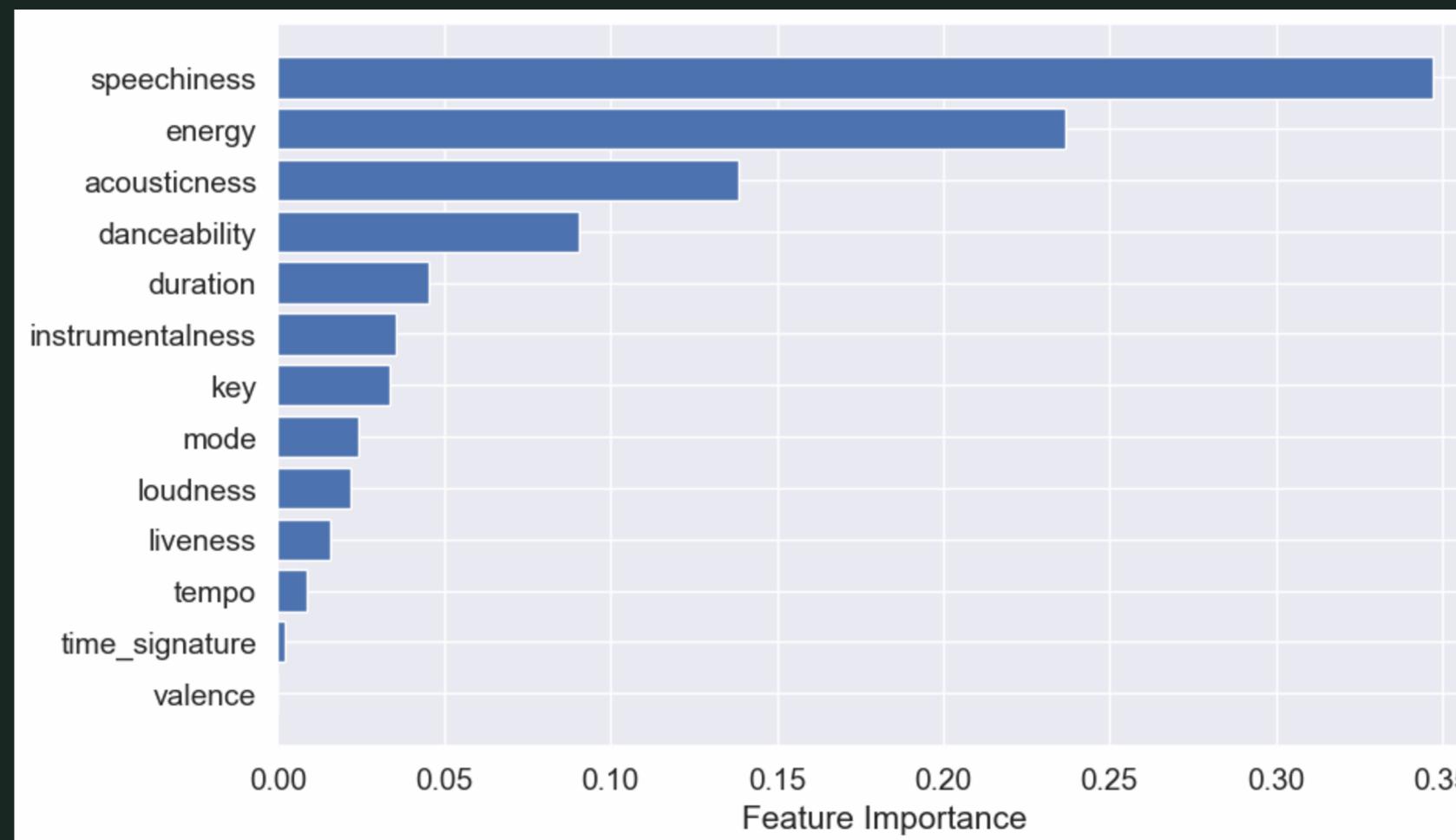
Escalamiento MinMax para “duration”, “tempo”, “loudness”



MODELOS

DECISION TREE CLASSIFIER

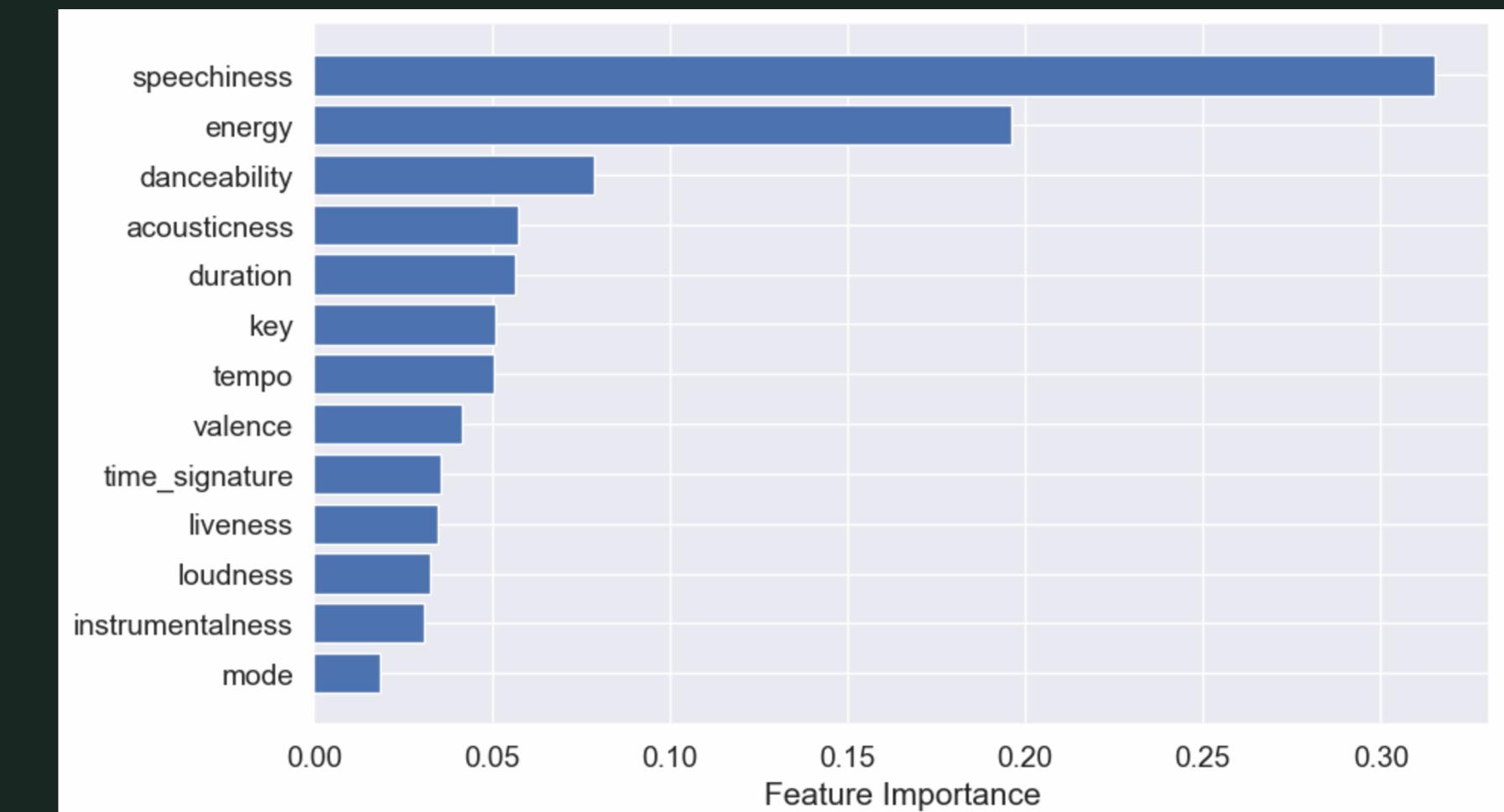
```
# Obtener la importancia de las características
importances = model.feature_importances_
feature_importance_df = pd.DataFrame({'Feature': x_train.columns, 'Importance': importances})
feature_importance_df = feature_importance_df.sort_values(by='Importance', ascending=True)
# Mostrar la importancia de las características
print("Importancia de las características:")
print(feature_importance_df)
```



XGB REGRESSOR

```
# Obtenemos la importancia de características
feature_importances = xgb_model.feature_importances_
feature_names = variables

# Ordenamos los índices
sorted_idx = feature_importances.argsort()
```



COLUMNAS INTERESANTES

Definidas en función de modelos presentados

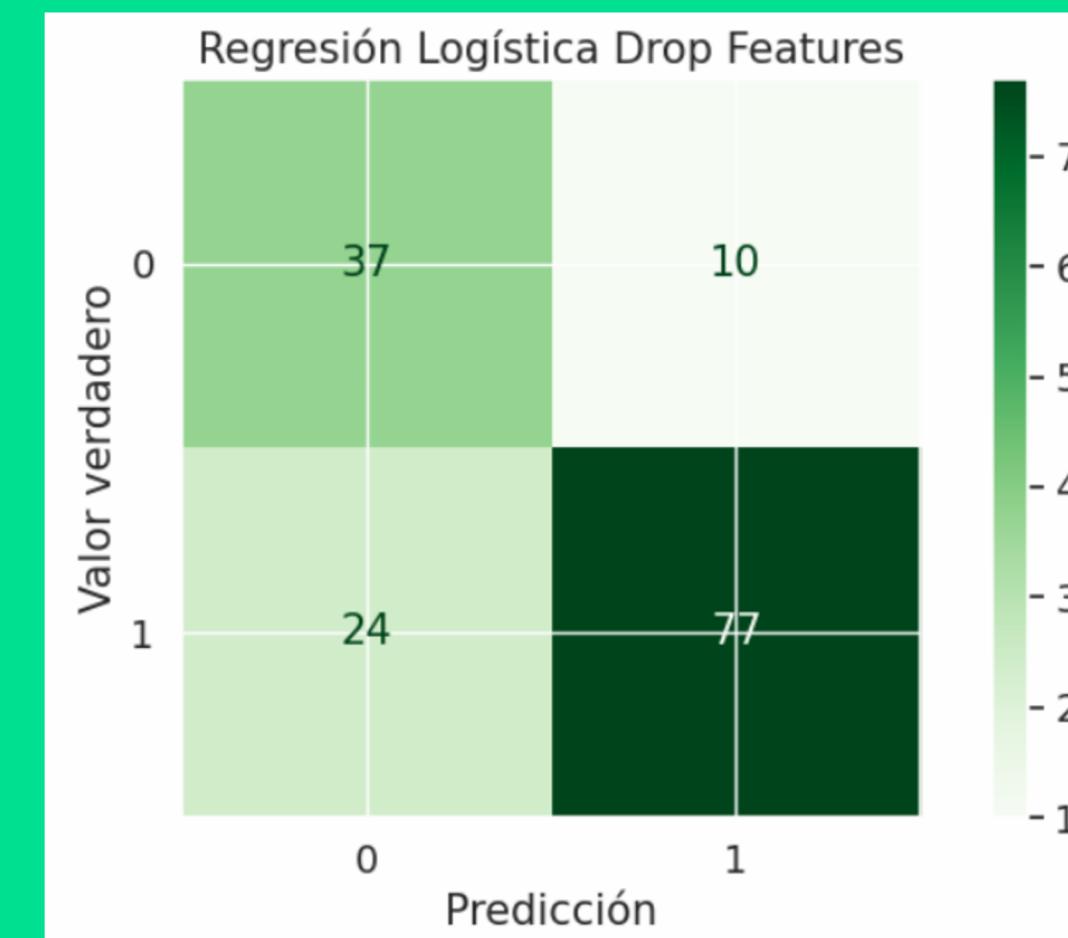
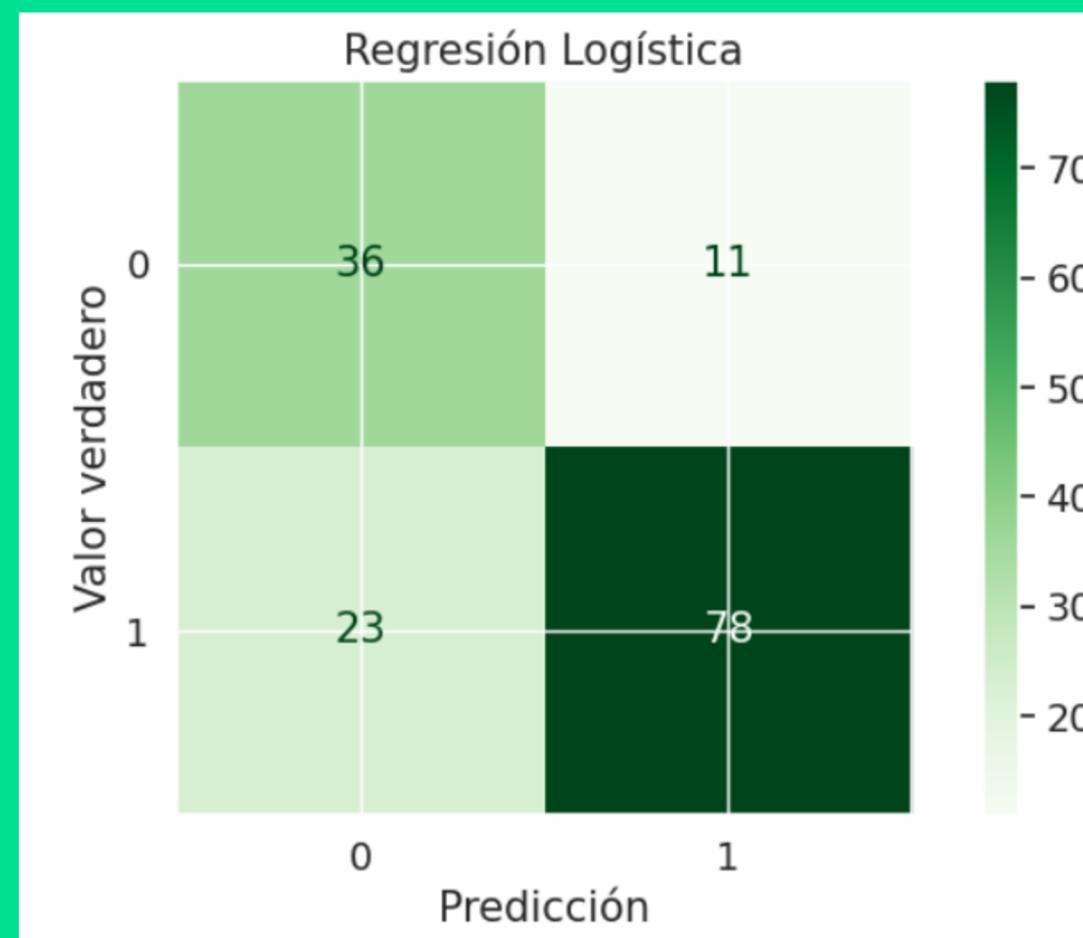
- SPEECHINESS
- ENERGY
- DANCEABILITY
- ACOUSTICNESS
- COLUMNA TARGET: LABEL

FEATURES NO RELEVANTES

- TIME_SIGNATURE
- MODE
- KEY
- LIVENESS
- TEMPO

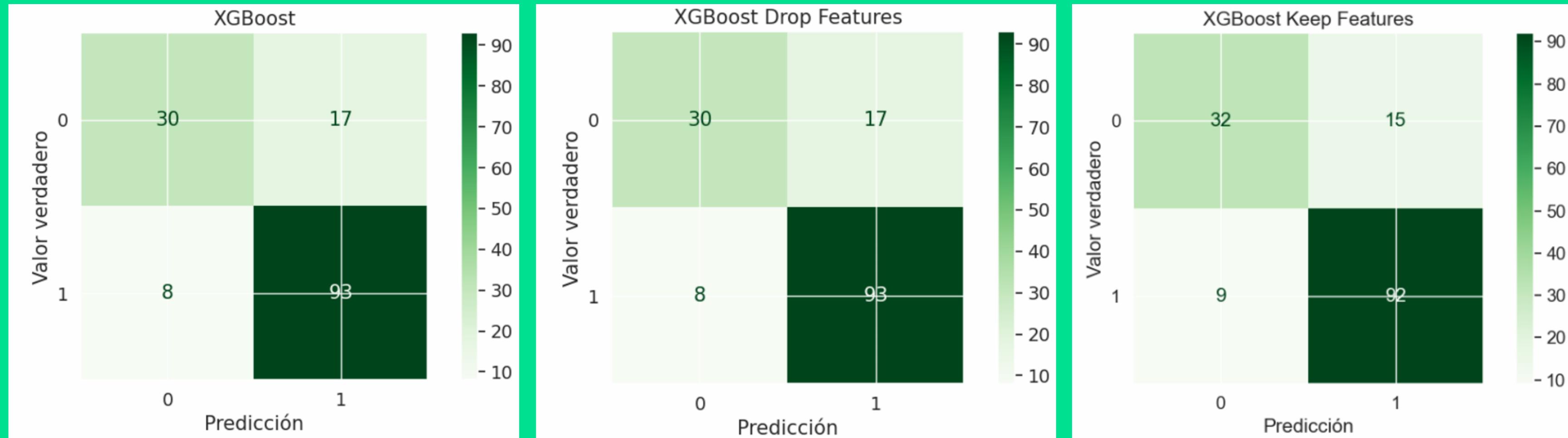
MODELOS PARA SELECCIÓN DE FEATURES

REGRESIÓN LOGÍSTICA



MODELOS PARA SELECCIÓN DE FEATURES

XGBOOST



DATASET FINAL

Definido únicamente a partir de los features más relevantes

```
1 keep_features = ['speechiness','energy','danceability','acousticness','label']
2 spotify_df_filter_scaler_feat_eng = spotify_df_filter_scaler[keep_features]
```

0.0s

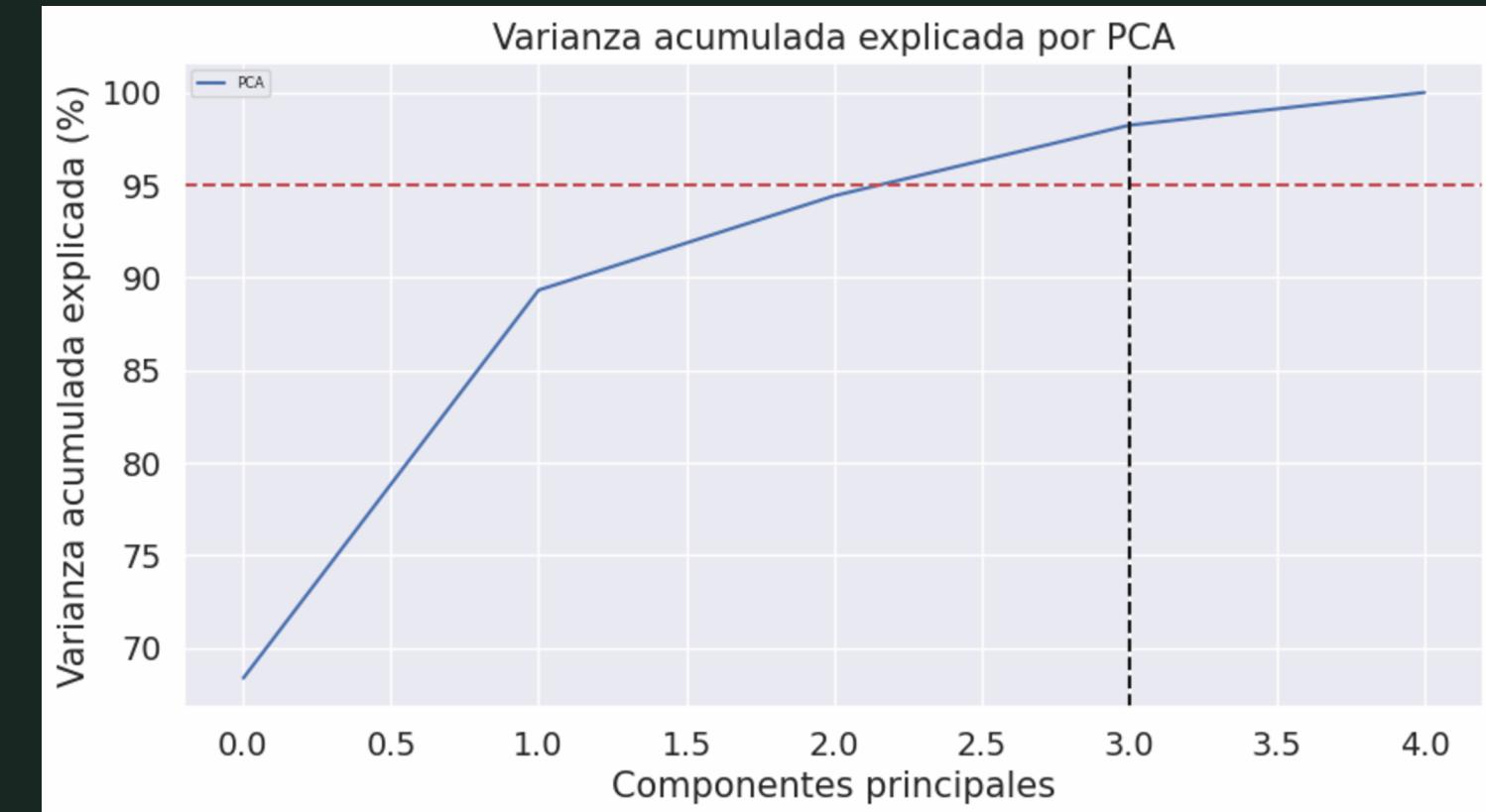
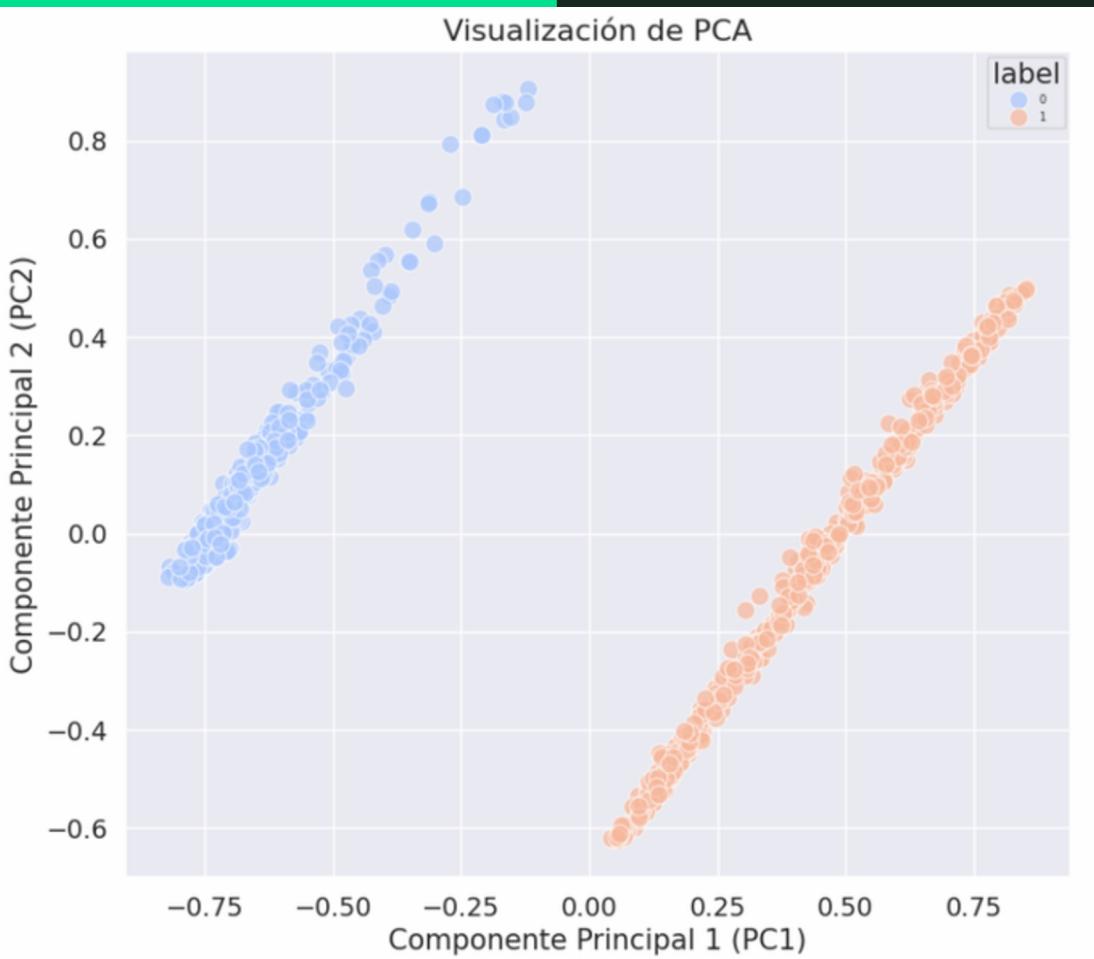
FEATURES MÁS
RELEVANTES

- SPEECHINESS
- ENERGY
- DANCEABILITY
- ACOUSTICNESS
- COLUMNA TARGET: LABEL

Reducción de la Dimensionalidad

PCA

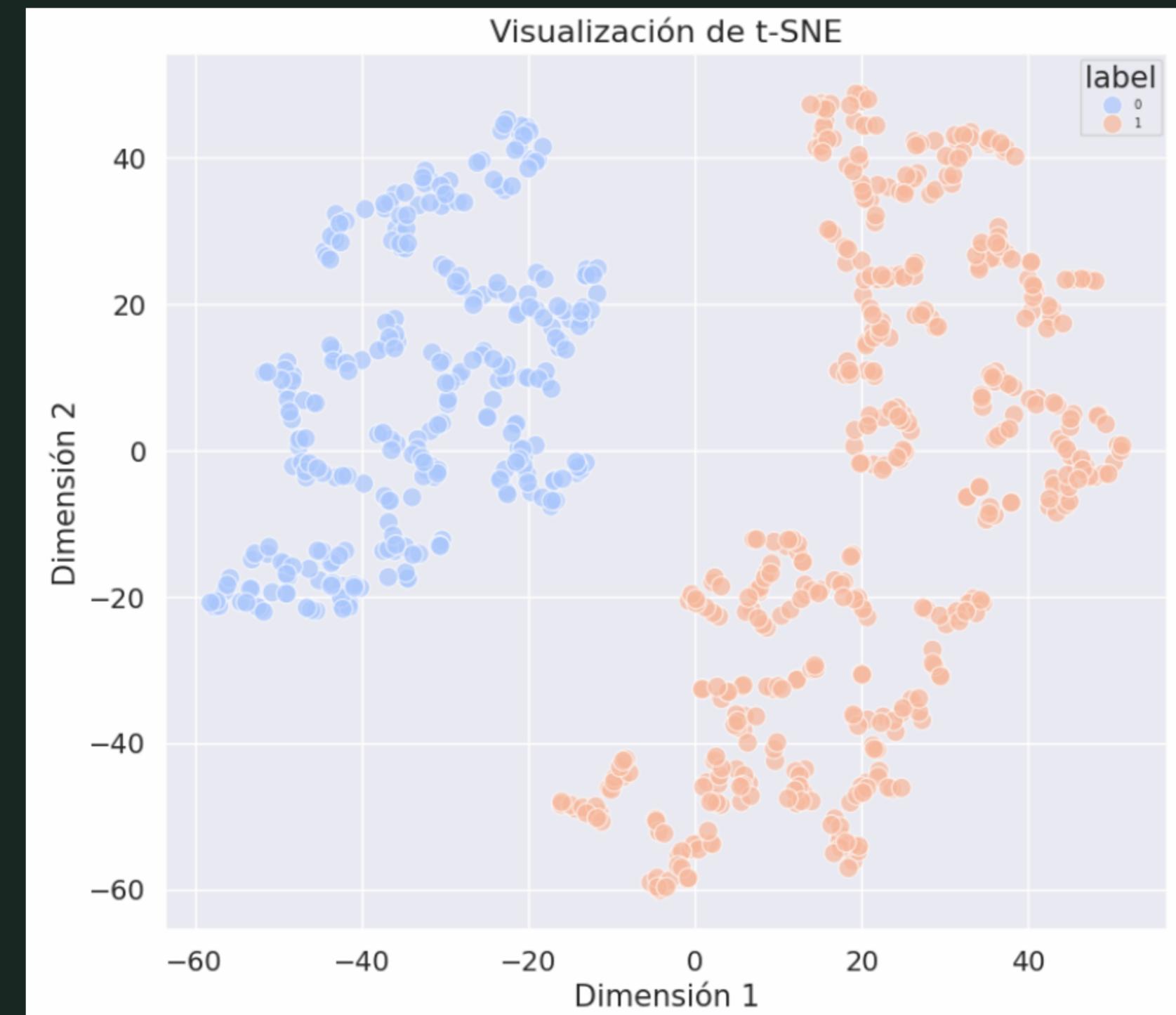
```
# Aplicar PCA
pca = PCA()
pca_projection = pca.fit_transform(spotify_df_filter_scaler_feat_eng)
```



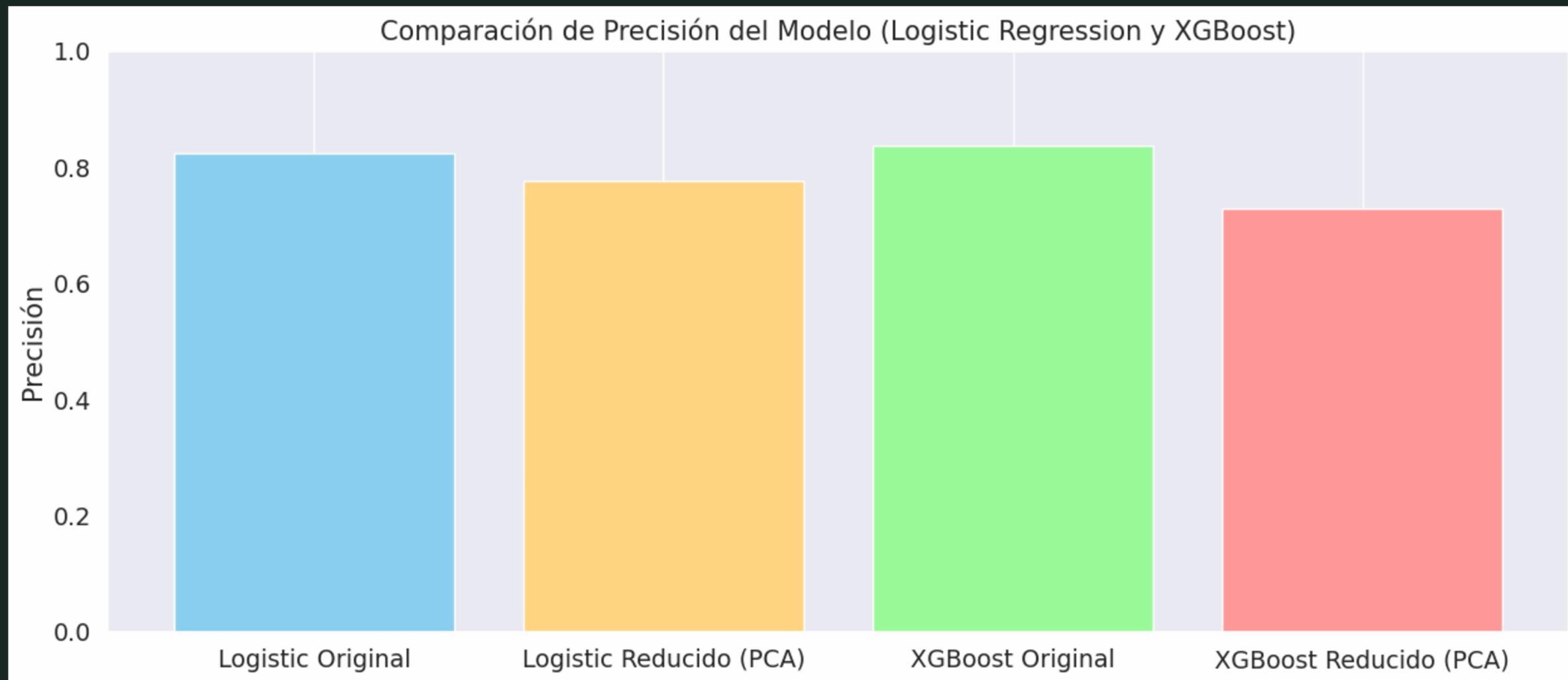
Reducción de la Dimensionalidad

t-SNE

```
# Aplicar t-SNE al dataset
tsne = TSNE(n_components=2, perplexity=10, random_state=20)
tsne_projection = tsne.fit_transform(spotify_df_filter_scaler_feat_eng)
```



Comparación de Precisión del Modelo (Logistic Regression y XGBoost)



¡Gracias!