Dataset de Spotify

Seminario 2

Jose Ramos

Universidad del Valle de Guatemala

Objetivos

1. Presentar el dataset
2. Intentar predecir la popularidad de una canción
3. Discutir recomendaciones

Dataset

1. Dos principales df: artistas, canciones.

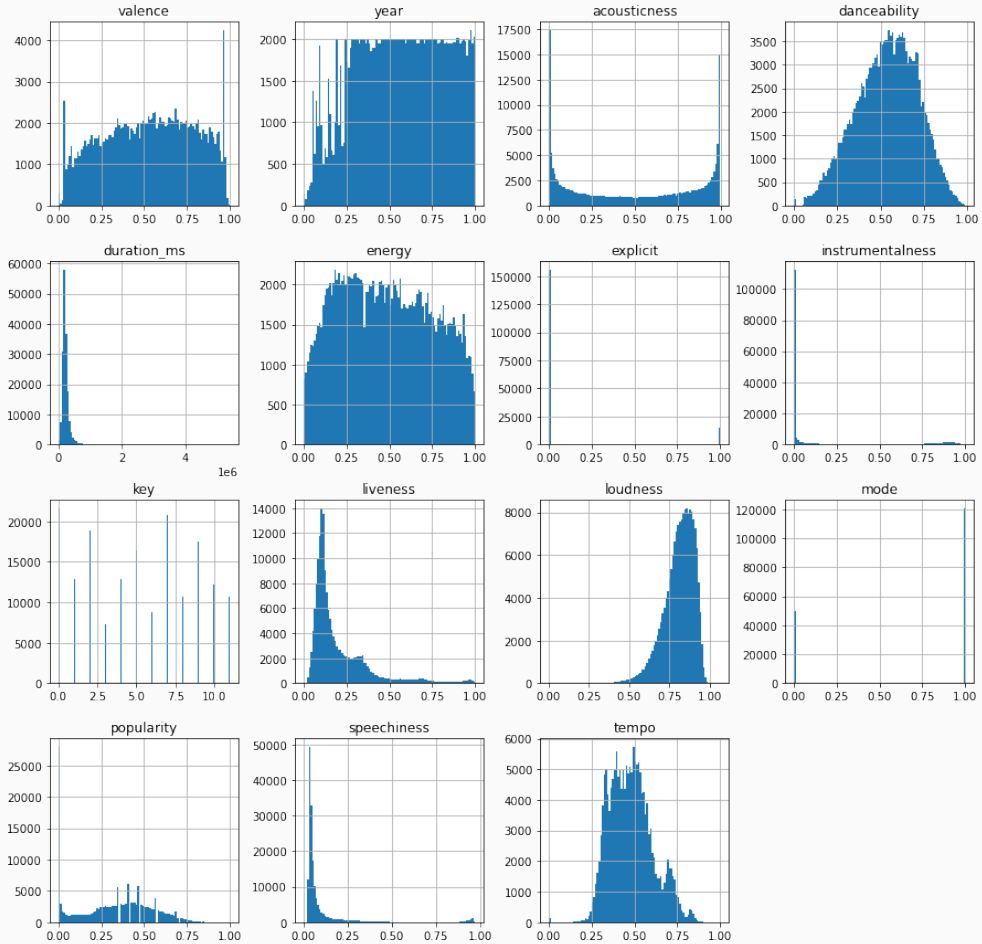


Figura 1: Histograma canciones
Fuente: elaboración propia

Dataset

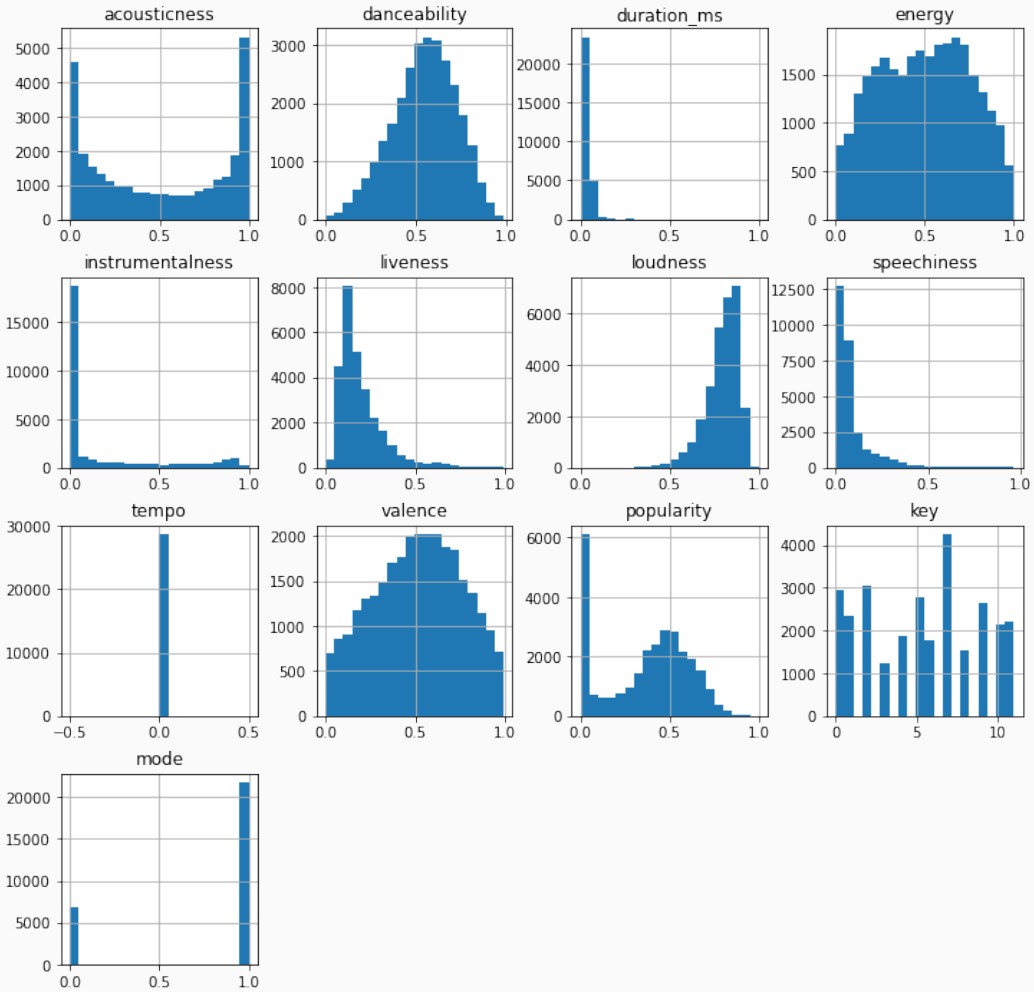


Figura 2: Histograma de los artistas
Fuente: elaboración propia

Visualización de datos

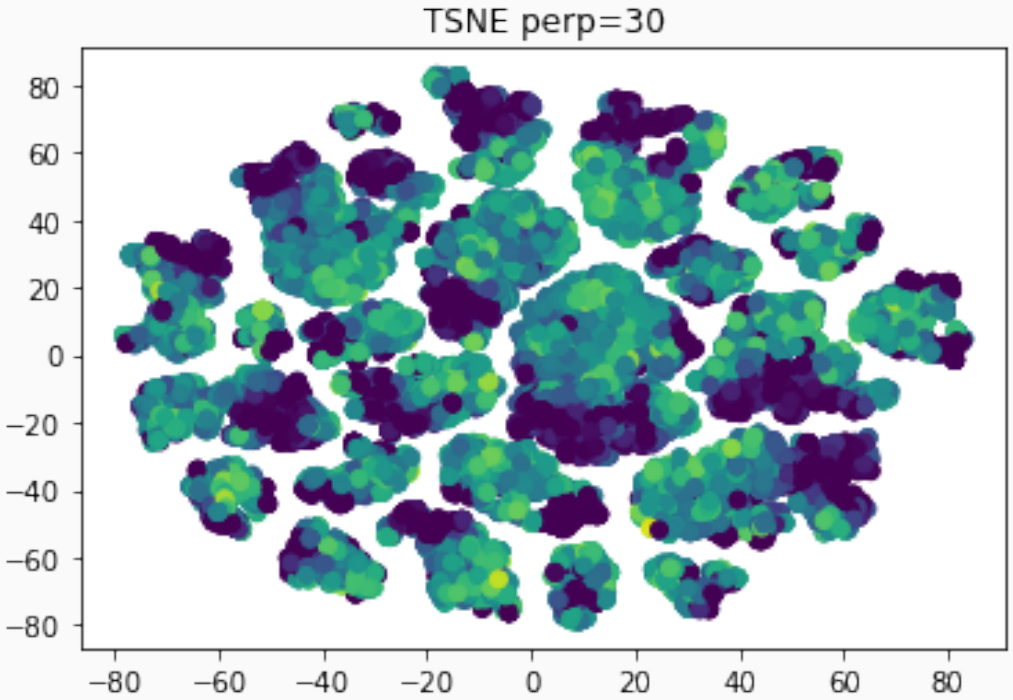


Figura 3: TSNE aplicado a una muestra de datos de las canciones usando la popularidad como el color
Fuente: elaboración propia

Prediciendo

Modelo inicial

1. Clasificar en 5 clases.
2. NO usar la popularidad del artista.

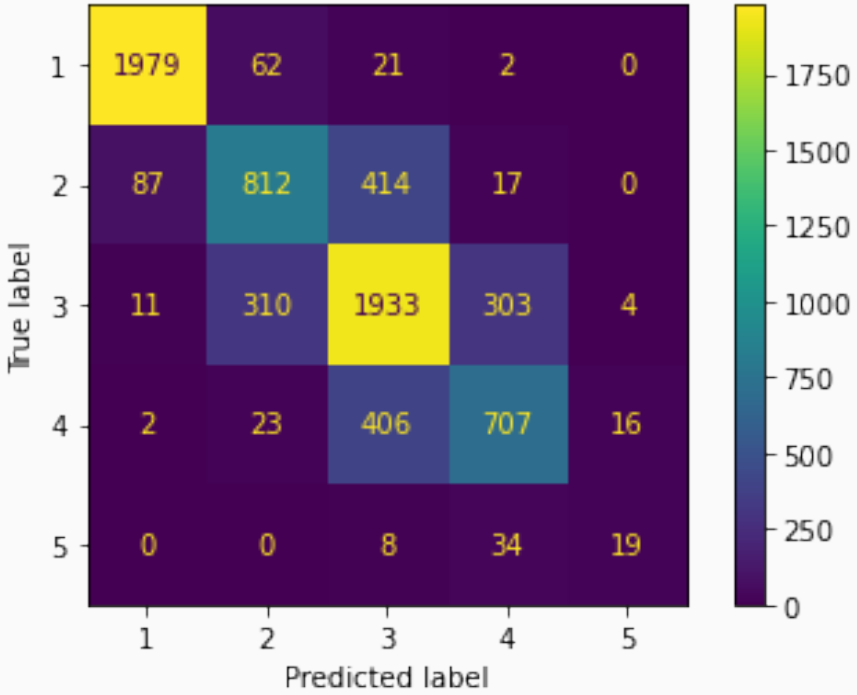


Figura 4: Matriz de confusión para 5 clases Random Forest
Fuente: elaboración propia

Modelo final

Usaremos solo dos clases, es decir convertimos el problema en uno de clasificación binaria.

Justificación: nos interesa saber si la canción va a ser popular (sí o no).

Modelo final

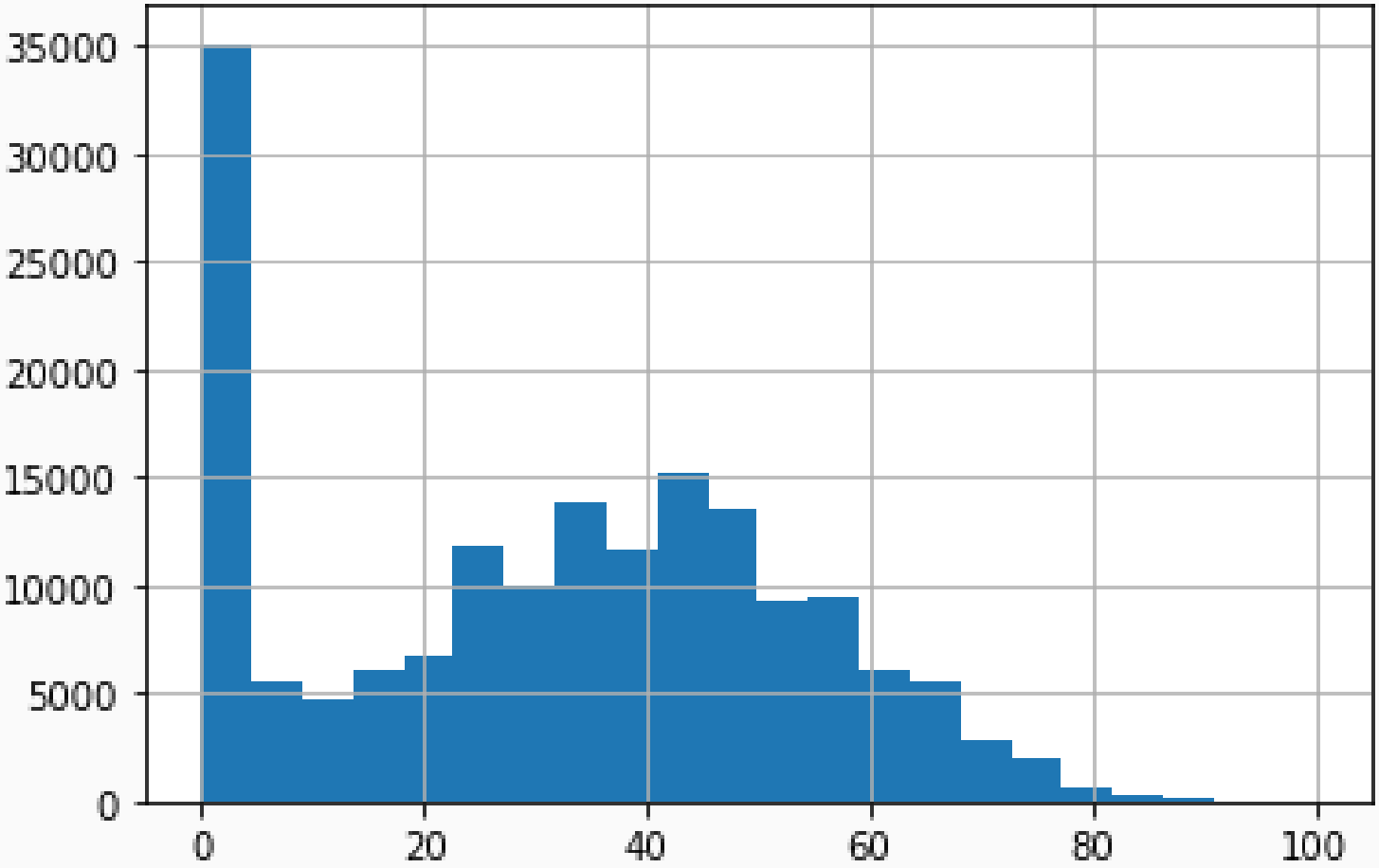
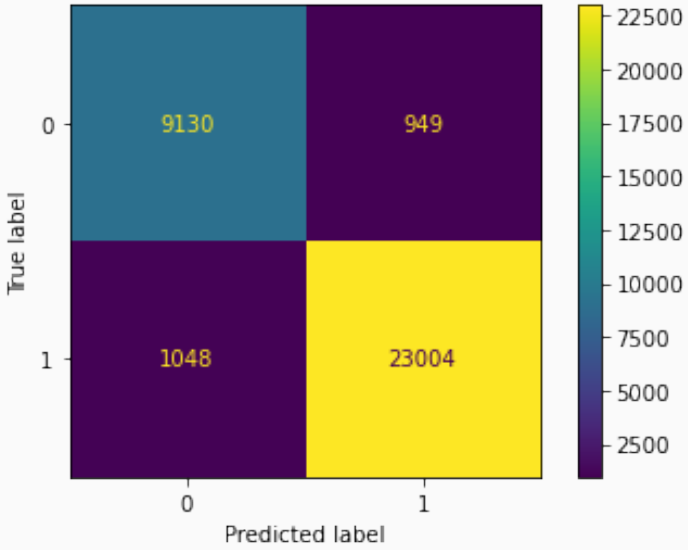
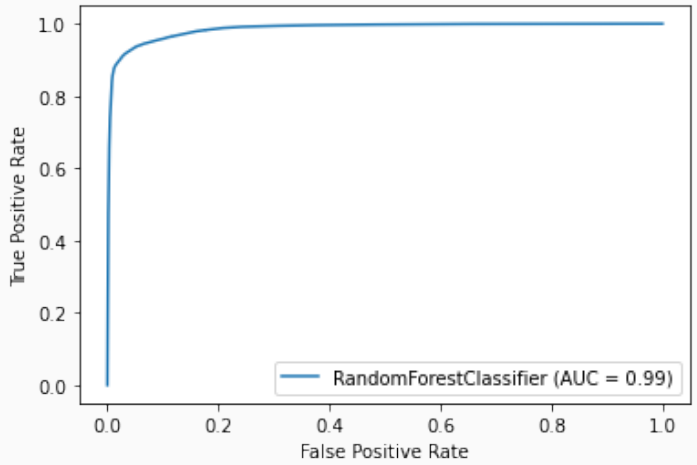


Figura 5: Histograma de las popularidades
Fuente: elaboración propia

Evaluación



(a) Matriz de confusión



(b) Área bajo la curva

Figura 6: Métricas para Random Forest
Fuente: elaboración propia

Variables importantes

```
Feature: valence, Score: 0.02193
Feature: year, Score: 0.53209
Feature: acousticness, Score: 0.14203
Feature: danceability, Score: 0.02244
Feature: duration_ms, Score: 0.03543
Feature: energy, Score: 0.06931
Feature: explicit, Score: 0.00247
Feature: instrumentalness, Score: 0.03301
Feature: key, Score: 0.01035
Feature: liveness, Score: 0.02200
Feature: loudness, Score: 0.04306
Feature: mode, Score: 0.00252
Feature: speechiness, Score: 0.04209
Feature: tempo, Score: 0.02127
```

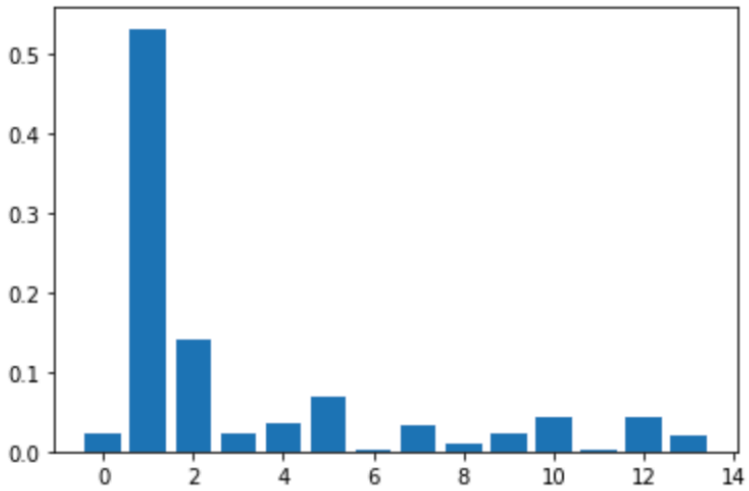
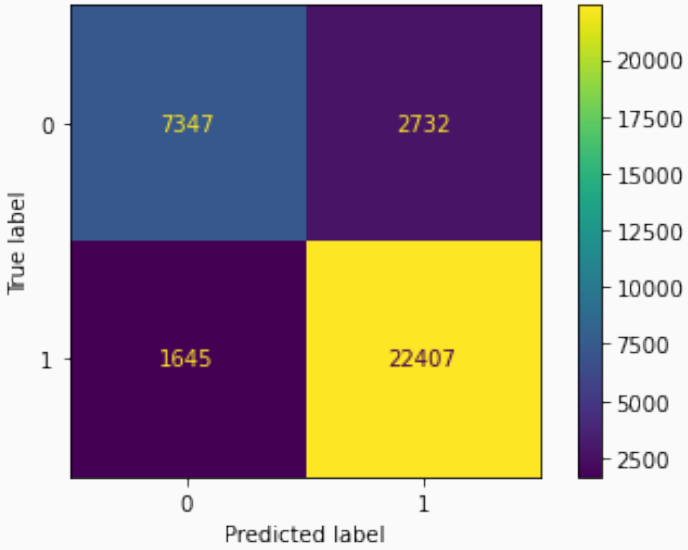
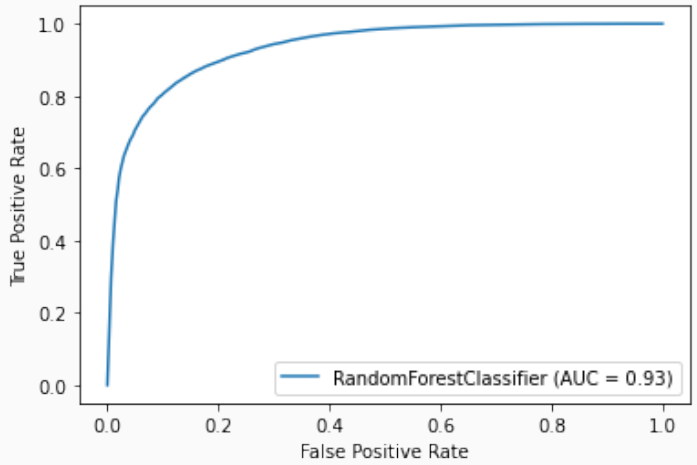


Figura 7: Atributos y sus pesos
Fuente: elaboración propia

Sin año



(a) Matriz de confusión



(b) Área bajo la curva

Figura 8: Métricas para Random Forest (sin año)
Fuente: elaboración propia

Nuevas variables importantes

```
Feature: valence, Score: 0.05562
Feature: acousticness, Score: 0.29541
Feature: danceability, Score: 0.05359
Feature: duration_ms, Score: 0.08337
Feature: energy, Score: 0.13291
Feature: explicit, Score: 0.00510
Feature: instrumentalness, Score: 0.06164
Feature: key, Score: 0.02418
Feature: liveness, Score: 0.05277
Feature: loudness, Score: 0.08480
Feature: mode, Score: 0.00602
Feature: speechiness, Score: 0.09738
Feature: tempo, Score: 0.04721
```

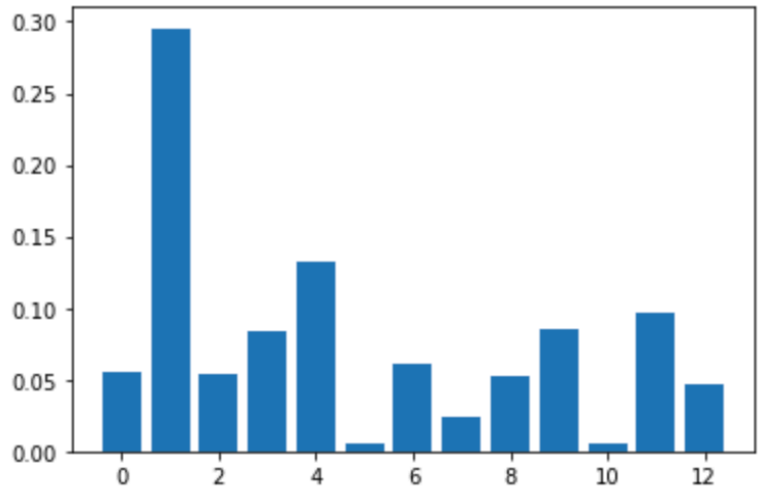


Figura 9: Atributos y sus pesos
Fuente: elaboración propia

Clustering

1. Ver si se podían agrupar por género.
2. Problema con la variable género.

```
In [141]: genres.unique(), len(genres.unique())

Out[141]: (array(["['show tunes']", '[]', "['comedy rock', 'comic', 'parody']", ...,
                  "['mainland chinese pop', 'zhongguo feng']",
                  "['c-pop', 'classic mandopop', 'mainland chinese pop', 'mandopop']",
                  "['chinese indie', 'chinese indie rock']"], dtype=object),
           10743)
```

Figura 10:
Fuente: elaboración propia

3. Recomendaciones para resolver el problema.
4. Clustering jerárquico
5. Word embeddings

Referencias

1. https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks