ITBA
Instituto Tecnológico de Buenos Aires

82.05 - Análisis Predictivo - Examen Final

60784 - Paula González

Contenidos

1. Introducción
2. Base de Datos
3. Analisis Exploratorio
4. Preprocesamiento de Datos
5. Modelos Predictivos
6. Posibles mejoras y conclusiones

Introducción

Caso de negocio

Predecir qué **contenidos son explícitos** en función a sus características para el desarrollo del segmento **Spotify Kids**.
El mismo implica:

- Entorno seguro y adecuado
- Contenido curado y apropiado para la edad

Base de Datos

Base de datos

El dataset analizado contiene canciones de Spotify en una variedad de **114 géneros** musicales. **Cada género incluye 1000 registros** e incluye las características sonoras del mismo.

A pesar de la divergencia que pueden llegar a tener los distintos géneros musicales, se optó por tomar la totalidad de los mismos.

Variables

Numéricas

- Speechiness [0-1]
- Acousticness [0-1]
- Instrumentalness [0-1]
- Liveness [0-1]
- Valence [0-1]
- Tempo
- Popularidad [0-100]
- Duración en milisegundos
- Danceability [0-1]
- Energy [0-1]
- Loudness (dB)

Categóricas

- Explicit T/F
- Key
- Genre
- Mode 0/1
- Time Signature

Extras

- Track_id
- Artista
- Album Name
- Track Name

Análisis Exploratorio

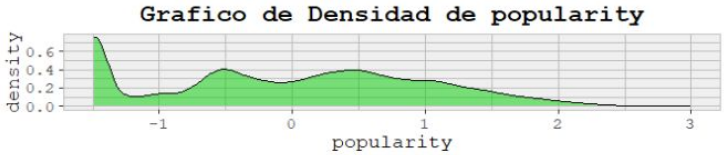


Grafico de Densidad de popularity
density
popularity

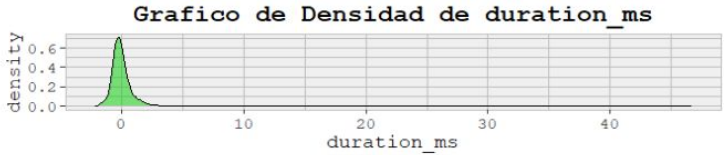


Grafico de Densidad de duration_ms
density
duration_ms

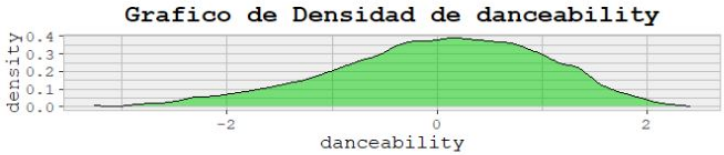


Grafico de Densidad de danceability
density
danceability

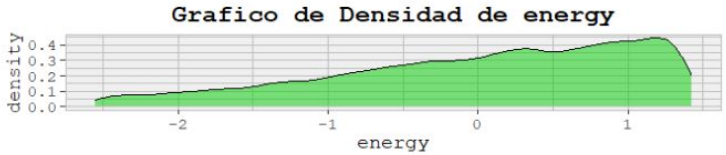


Grafico de Densidad de energy
density
energy

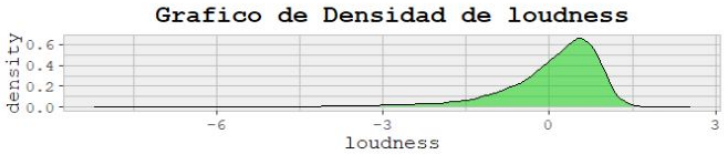


Grafico de Densidad de loudness
density
loudness


Grafico de Densidad de speechiness
density
speechiness

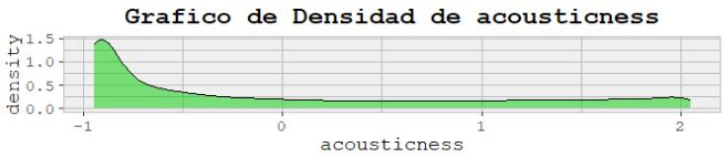


Grafico de Densidad de acousticness
density
acousticness

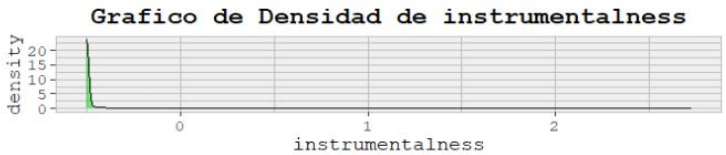


Grafico de Densidad de instrumentalness
density
instrumentalness

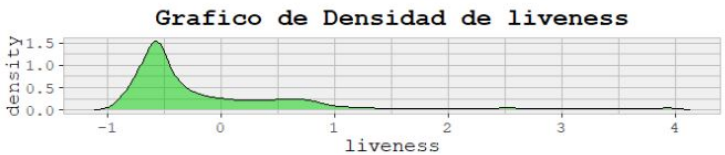


Grafico de Densidad de liveness
density
liveness

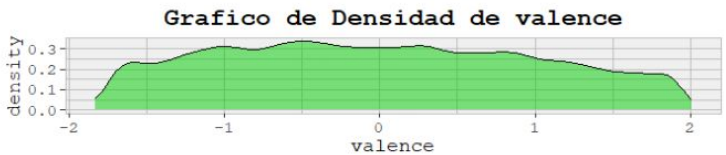
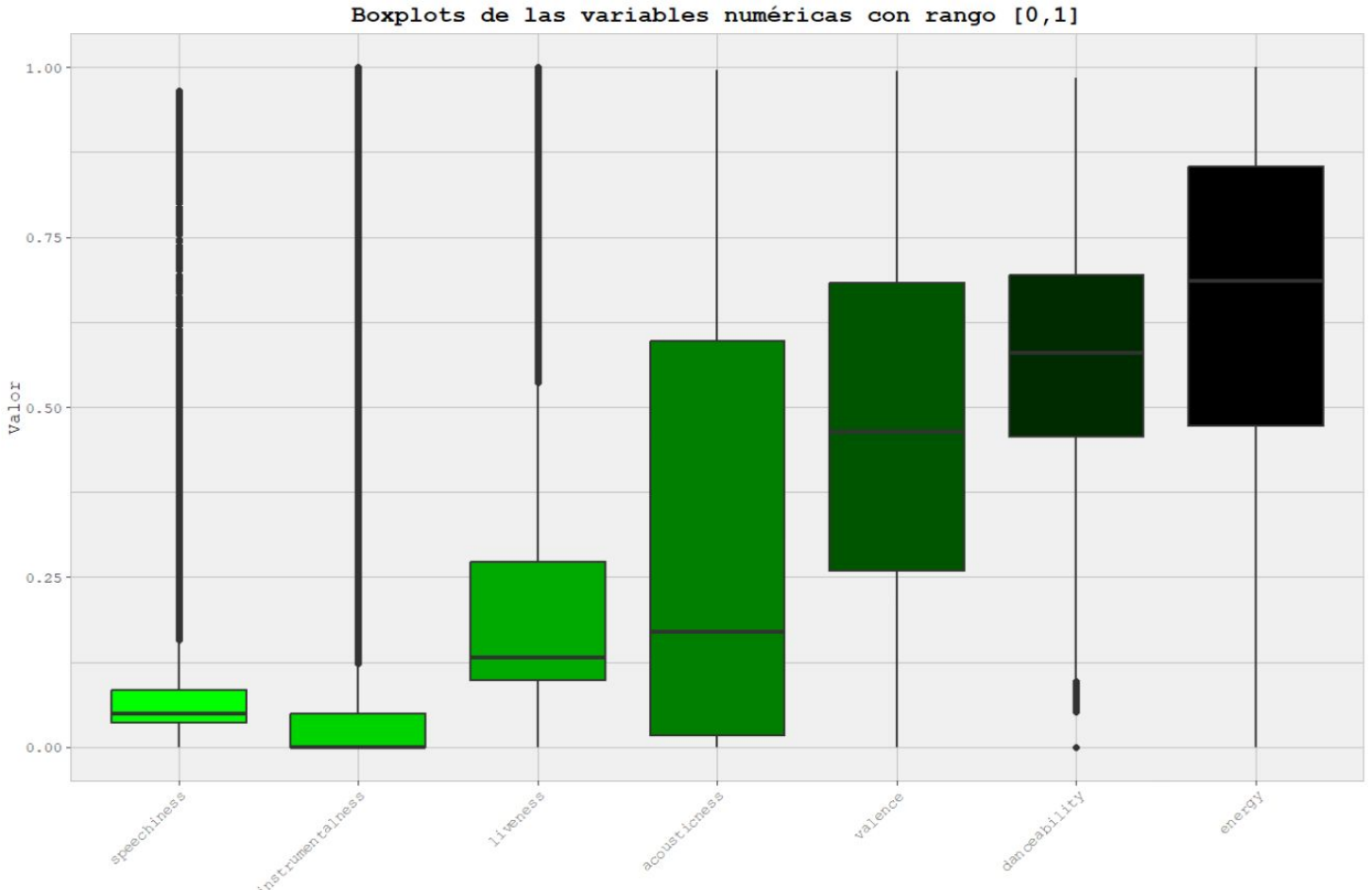


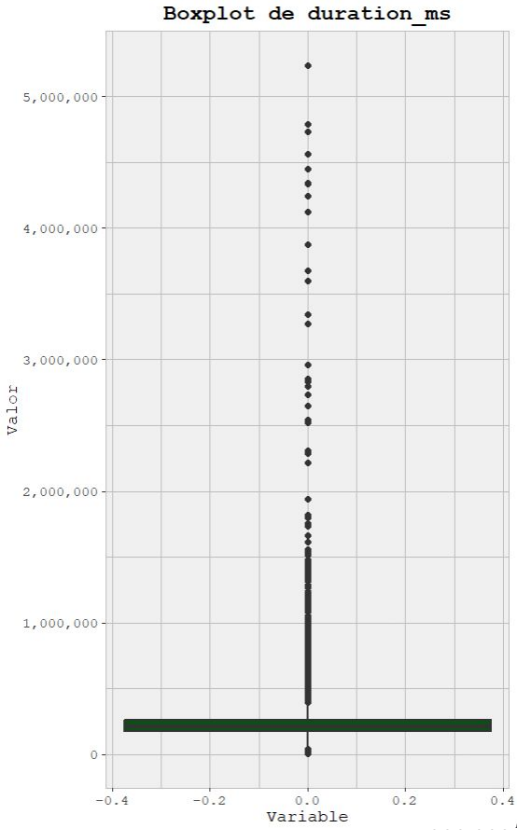
Grafico de Densidad de valence
density
valence

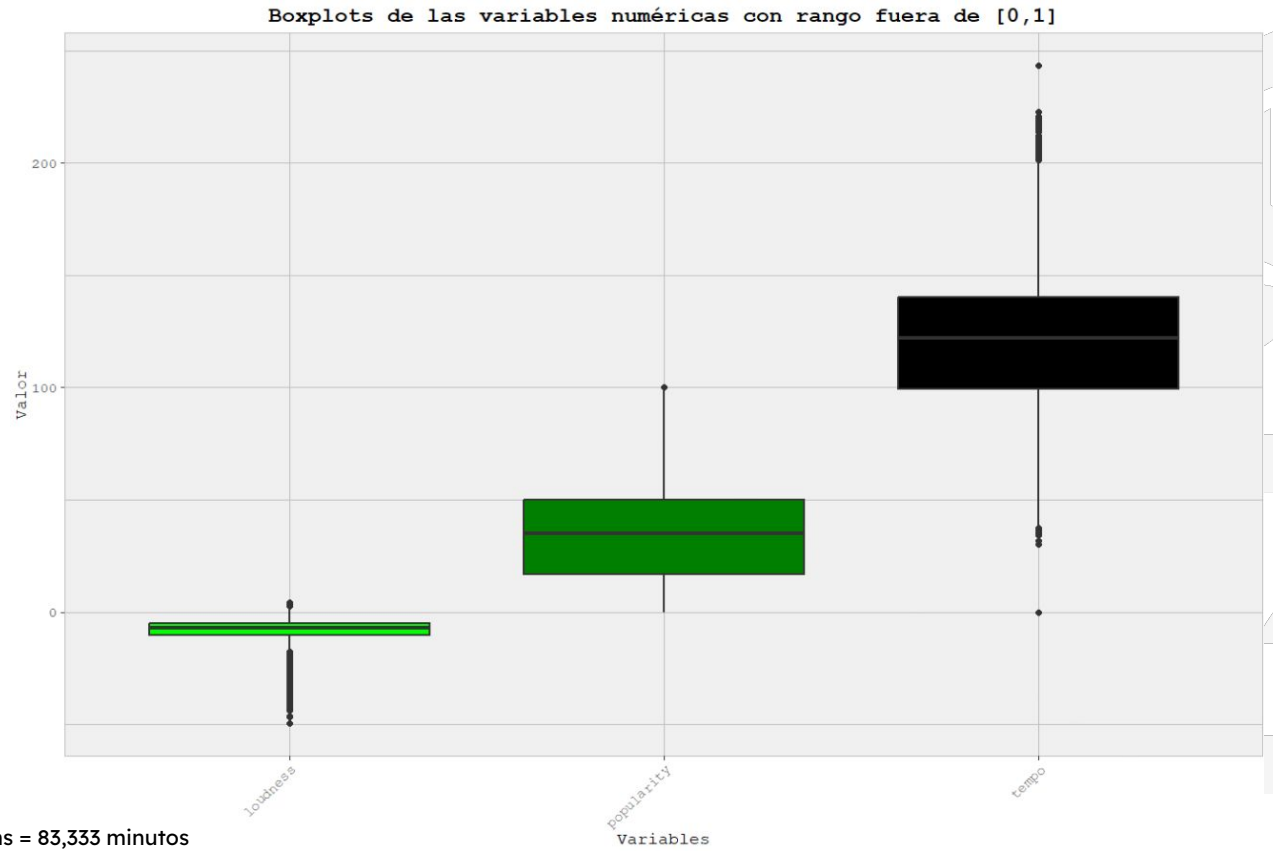

Grafico de Densidad de tempo
density
tempo

Outliers

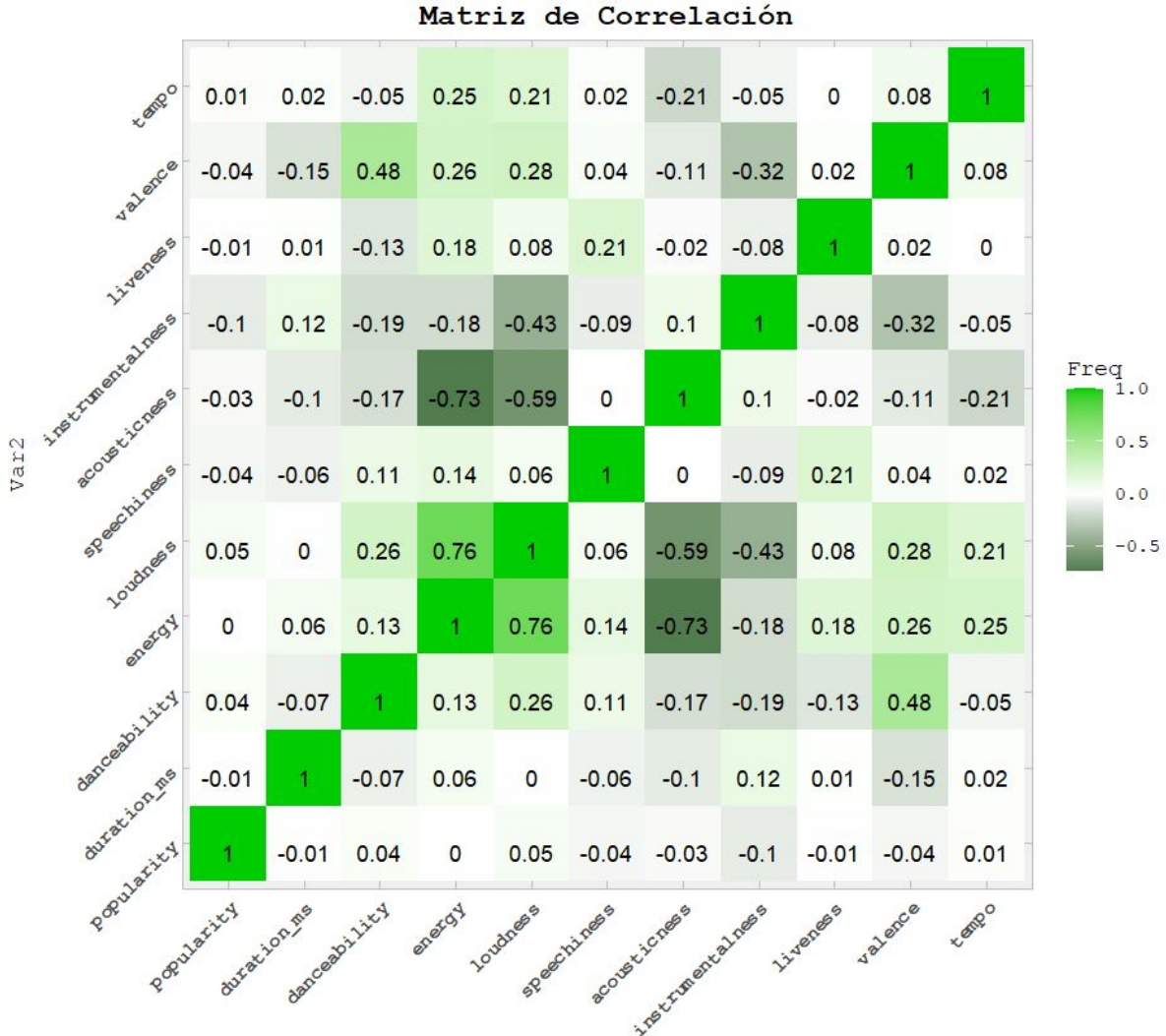

Boxplots de las variables numéricas con rango [0,1]
Valor
1.00
0.75
0.50
0.25
0.00
speechiness
instrumentalness
liveness
acousticness
valence
danceability
energy

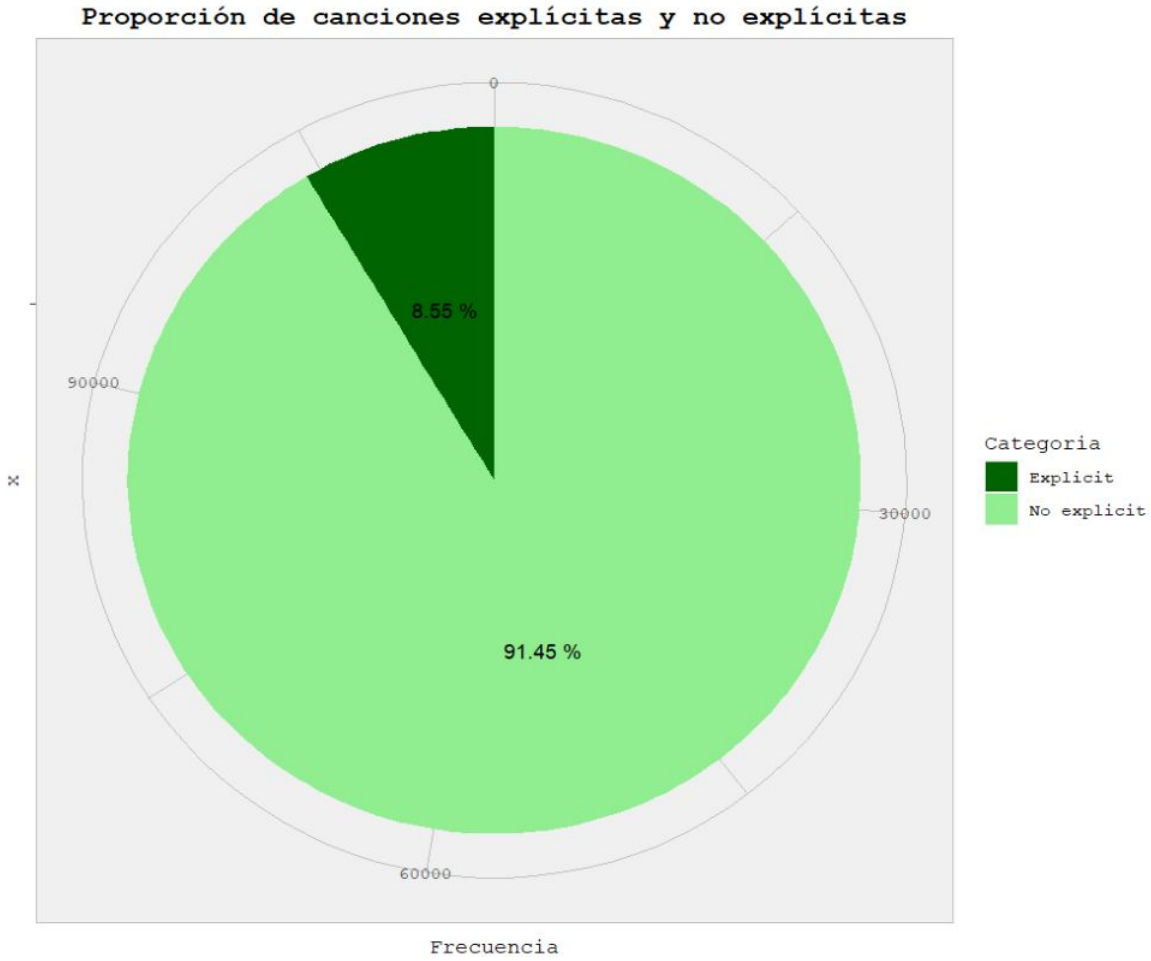
Outliers



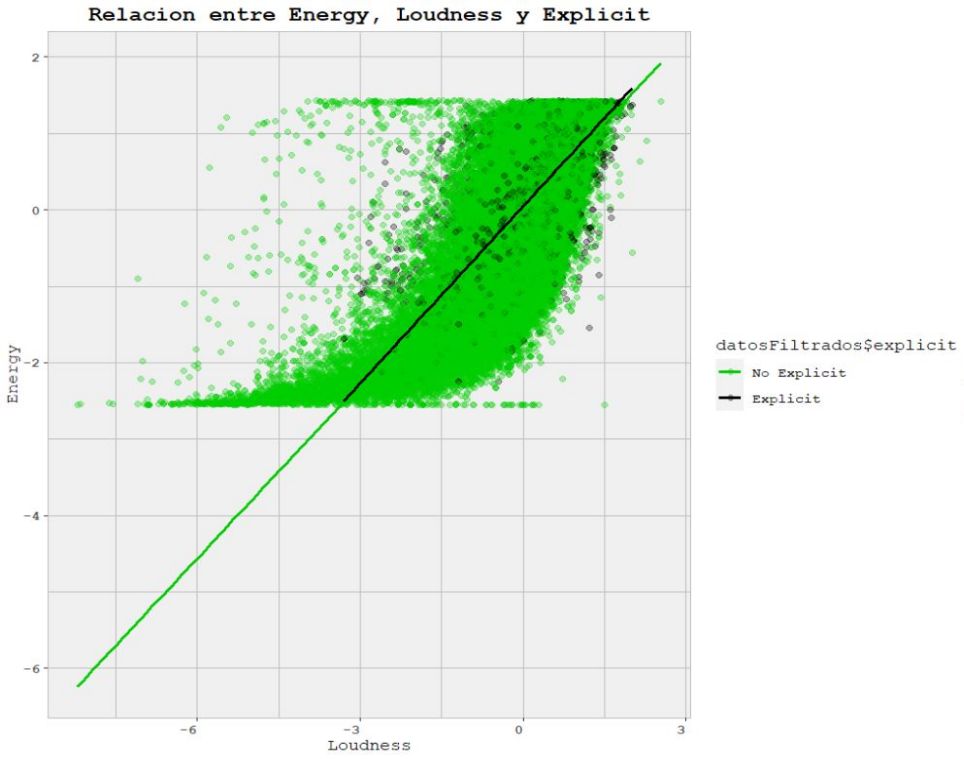


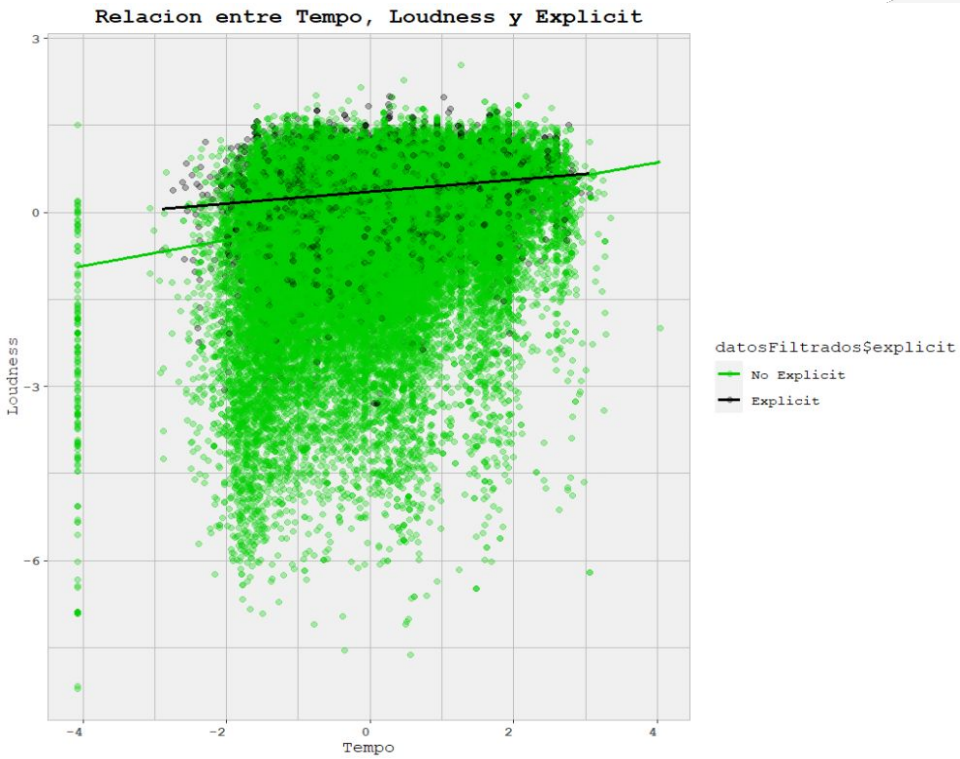
5,000,000ms = 83,333 minutos


Matriz de Correlación
Var2
tempo
valence
liveness
instrumentalness
acousticness
speechiness
loudness
energy
danceability
duration_ms
popularity
0.01 0.02 -0.05 0.25 0.21 0.02 -0.21 -0.05 0 0.08 1
-0.04 -0.15 0.48 0.26 0.28 0.04 -0.11 -0.32 0.02 1 0.08
-0.01 0.01 -0.13 0.18 0.08 0.21 -0.02 -0.08 1 0.02 0
-0.1 0.12 -0.19 -0.18 -0.43 -0.09 0.1 1 -0.08 -0.32 -0.05
-0.03 -0.1 -0.17 -0.73 -0.59 0 1 0.1 -0.02 -0.11 -0.21
-0.04 -0.06 0.11 0.14 0.06 1 0 -0.09 0.21 0.04 0.02
0.05 0 0.26 0.76 1 0.06 -0.59 -0.43 0.08 0.28 0.21
0 0.06 0.13 1 0.76 0.14 -0.73 -0.18 0.18 0.26 0.25
0.04 -0.07 1 0.13 0.26 0.11 -0.17 -0.19 -0.13 0.48 -0.05
-0.01 1 -0.07 0.06 0 -0.06 -0.1 0.12 0.01 -0.15 0.02
1 -0.01 0.04 0 0.05 -0.04 -0.03 -0.1 -0.01 -0.04 0.01
popularity
duration_ms
danceability
energy
loudness
speechiness
acousticness
instrumentalness
liveness
valence
tempo
Freq
1.0
0.5
0.0
-0.5


Proporción de canciones explícitas y no explícitas
0
30000
60000
90000
8.55 %
91.45 %
Categoria
Explicit
No explicit
Frecuencia

Explicit en relación de otras variables





Asociación entre Explicit y las categóricas

Variable1	Variable2	V-Cramer
explicit	key	0.068
	mode	0.037
	time_signature	0.06
	genre	0.402

Asociación entre Explicit y las numéricas

Variable1	Variable2	p-valor
explicit	energy	≈ 0
	tempo	0.35
	loudness	≈ 0
	popularity	≈ 0
	speechiness	≈ 0
	duration_ms	≈ 0
	liveness	≈ 0
	valence	0.201
	acousticness	≈ 0
	danceability	≈ 0
	instrumentalness	≈ 0

Preprocesamiento

Preprocesamiento de datos

Se realizaron **modificaciones** en la base de datos original, incluyendo la eliminación y adición de columnas, además de tratamientos como **limpieza**, **imputación** de valores **faltantes** y generación de nuevas características, con el fin de preparar los datos para análisis posteriores.

- **Escalado** de variables numéricas (en azul)
- **Label encoding** para algunas variables categóricas no-numéricas (en verde)

Numéricas

- Speechiness [0-1]
- Acousticness [0-1]
- Instrumentalness [0-1]
- Liveness [0-1]
- Valence [0-1]
- Tempo
- Popularidad [0-100]
- Duración en milisegundos
- Danceability [0-1]
- Energy [0-1]
- Loudness (dB)

Categóricas

- Explicit ~~T/F~~ → 1/0
- Key
- Genre
- Mode 0/1
- Time Signature

Extras

- ~~Track_id~~
- Artista
- ~~Album Name~~
- ~~Track Name~~

Columnas adicionales

- Valor **mínimo, máximo, mediana y promedio** de las siguientes características numéricas **por género**:
  - Energy, Danceability, Instrumentalness, Speechiness, Acousticness.
- Valor **mínimo, máximo y mediana** de la **duración** de las canciones **por artista**
- **Promedio** de la cantidad de canciones **explicit por género y por artista**.

Base definitiva para el modelo

Al finalizar todo el preprocesamiento, obtuve una tabla de **42 columnas y 114000 registros**. Las columnas son las que se notan a continuación.

```
data.columns
```

✓ 0.0s

```
Index(['popularity', 'duration_ms', 'explicit', 'danceability', 'energy',
       'key', 'loudness', 'mode', 'speechiness', 'acousticness',
       'instrumentalness', 'liveness', 'valence', 'tempo', 'time_signature',
       'energy_max_X_track_genre', 'energy_min_X_track_genre',
       'energy_median_X_track_genre', 'energy_mean_track_genre',
       'danceability_max_X_track_genre', 'danceability_min_X_track_genre',
       'danceability_median_X_track_genre', 'danceability_mean_track_genre',
       'instrumentalness_max_X_track_genre',
       'instrumentalness_min_X_track_genre',
       'instrumentalness_median_X_track_genre',
       'instrumentalness_mean_track_genre', 'speechiness_max_X_track_genre',
       'speechiness_min_X_track_genre', 'speechiness_median_X_track_genre',
       'speechiness_mean_track_genre', 'acousticness_max_X_track_genre',
       'acousticness_min_X_track_genre', 'acousticness_median_X_track_genre',
       'acousticness_mean_track_genre', 'max_duration_by_artist',
       'min_duration_by_artist', 'median_duration_by_artist',
       'promedio_explicit_por_artista', 'promedio_explicit_por_genero',
       'artist_encoded', 'genre_encoded'],
      dtype='object')
```

Modelos predictivos

Variables

- **Target** (y): “explicit”
- **Predictora** (X): las 41 variables

Métricas de evaluación

1. Exactitud

$$Exactitud = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones}$$

2. Precisión

$$Precisión = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ positivos}$$

3. Recall (sensibilidad)

$$Recall = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ negativos}$$

4. AUC-ROC

Área bajo la curva ROC dice qué tan bueno es el modelo para distinguir entre explicit y no explicit.

Probando modelos

Con un test size del 0.2.

Modelo	Exactitud	Precisión	Recall	AUC-ROC
Regresión Logística	0.925	0.621	0.285	0.881
XGBOOST	0.974	0.896	0.783	0.995
Decision Tree	0.974	0.896	0.783	0.989
Random Forest	0.974	0.896	0.783	0.993

Hiperparametros del RandomForest → n_estimators=2000, max_depth=None, min_samples_split=2, min_samples_leaf=1, random_state=42

Conclusiones

Posibles mejoras

- Realizar un **feature importance** en el modelo ganador
- Realizar un **gráfico de curvas AUC-ROC** para mejorar la visualización de los resultados.
- **Ampliar la información** acerca de las canciones que tengan que ver con métricas sobre los usuarios que escuchan las canciones.

Conclusiones

- El tempo y loudness es mayor en las canciones explícitas que en las no-explícitas.
- Como sólo un 8,55% de las canciones son explícitas no habría que filtrar tantas canciones para el armado del Spotify Kids.
- El tempo, la valence y el género de una canción influye significativamente en si una canción es explícita o no.
- Teniendo en cuenta las métricas elegidas, RandomForest es el mejor modelo para predecir las canciones explícitas.

ITBA
Instituto Tecnológico de Buenos Aires

¡Muchas gracias!

60784 - Paula González

paulgonzalez@itba.edu.ar