ANÁLISIS PREDICTIVO DE LA POPULARIDAD DE LAS CANCIONES

RESUMEN EJECUTIVO - PARTE I - GRUPO 6

GITHUB

ÍNDICE

EL EQUIPO

Alejandro Zielke
Estadística - LMU München

Iván Gálvez
Estadística - UPC/UB

Margherita Berutti
Data Science - UniTn

Alejandro Carnero
Estadística - UPC/UB

Diego Carrasco
Estadística - UPC/UB

UNIVERSITAT DE BARCELONA

UPC

ENTENDIMIENTO DE NEGOCIO Y DATOS

Objetivos y enfoque del proyecto

OBJETIVO PRINCIPAL

Crear el modelo predictivo más preciso posible

El KPI utilizado para medir este objetivo será el Mean Absolute Percentage Error (MAPE). Cuanto más bajo, mejor.

OBJETIVOS SECUNDARIOS

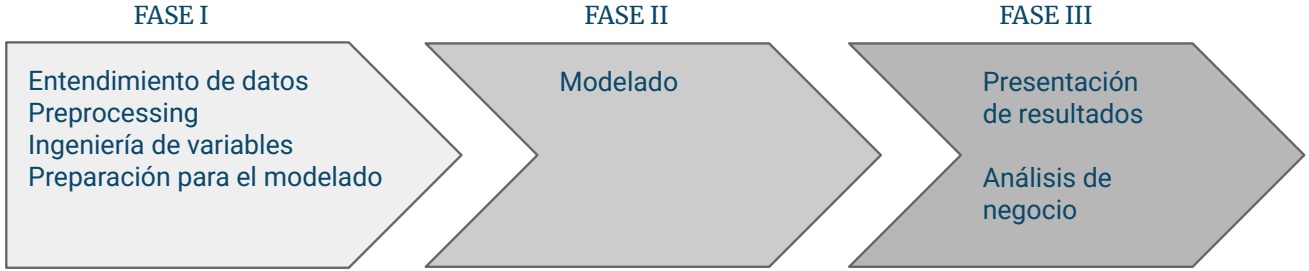
Entender los datos

Describir el comportamiento de las variables en nuestra base de datos

Canción óptima

Qué características debe tener una canción para ser popular

Fases del proyecto



ENTENDIMIENTO DE DATOS

Variable	Descripción	Tipología	Rango / Unidades	Rol	Comentarios
song_popularity	Puntuación continua de popularidad.	**Numérica (Entera)**	[0, 100]	**Objectiu (Target)**	Relativamente simétrica (Skew -0.51).
ID	Identificador único de la canción.	**Numérica (ID)**	N/A	**Identificador**	No usar para modelar.
danceability	Mide qué tan adecuada es la canción para bailar.	**Numérica (Continua)**	[0, 1]	Explicativa	Simétrica (Skew -0.38).
energy	Intensidad y nivel de actividad de la canción.	**Numérica (Continua)**	[0, 1]	Explicativa	Simétrica (Skew -0.60).
acousticness	Grado en que la pista es acústica.	**Numérica (Continua)**	[0, 1]	Explicativa	Asimetría positiva (Skew 1.08).
instrumentalness	Probabilidad de que no contenga voces.	**Numérica (Continua)**	[0, 1]	Explicativa	**Alta asimetría positiva** (Skew 3.09). Exceso de ceros.
liveness	Probabilidad de que la pista fuera grabada en vivo.	**Numérica (Continua)**	[0, 1]	Explicativa	**Alta asimetría positiva** (Skew 2.25).

ENTENDIMIENTO DE DATOS

Variables	Descripción	Tipología	Rango / Unidades	Rol	Comentarios (Calidad de Datos)
audio_valence	Positividad musical (feliz $\leftrightarrow$ triste).	**Numérica (Continua)**	[0, 1]	Explicativa	Muy simétrica. Distribución plana.
tempo	Velocidad de la canción.	**Numérica (Continua)**	BPM	Explicativa	Simétrica (Skew 0.42).
loudness	Volumen medio.	**Numérica (Continua)**	dB (Negativos)	Explicativa	**Asimetría negativa** (Skew -1.89).
speechiness	Presencia de palabras habladas.	**Numérica (Continua)**	[0, 1]	Explicativa	**Alta asimetría positiva** (Skew 2.26).
song_duration_ms	Duración total de la canción.	**Numérica (Entera)**	Milisegundos	Explicativa	**Asimetría extrema** (Skew 4.03). Outliers.
key	Tonalidad musical (0=C, ..., 11=B).	**Categórica Nominal** (Factor, 12 niveles)	[0-11]	Explicativa	Se transformó a factor por su naturaleza.
audio_mode	Modalidad musical (0=Menor, 1=Mayor).	**Categórica Binaria** (Factor, 2 niveles)	[0, 1]	Explicativa	Se transformó a factor por su naturaleza.
time_signature	Cantidad de notas por compás.	**Categórica Nominal** (Factor, 5 niveles)	[0, 1, 3, 4, 5]	Explicativa	Se transformó a factor por su naturaleza.

UPC
UNIVERSITAT DE BARCELONA

FRAMEWORK METODOLÓGICO

Proceso	Margherita Berutti	Alejandro Zielke	Alejandro Carnero	Iván Gálvez	Diego Carrasco
EDA	■	■	■		
Missings				■	
Outliers					■
Ingeniería de variables			■		
Reglas de asociación	■				
Modelado		■			
Presentación de resultados			■		■

ANÁLISI DE DATOS EXPLORATORIO

VARIABLE OBJETIVO

La variable *popularity* (numérica) es simétrica y no está sesgada (Media 44.4 vs. Mediana 45.0).

POSIBLES PREDICTORES

- **Positivos**: *danceability* y *energy* son los predictores positivos más fuertes. A mayor popularidad, mayor es su mediana.
- **Negativos**: *acousticness* e *instrumentalness* son los predictores negativos más fuertes. A mayor popularidad, menor es su mediana.
- **Débiles**: *liveness*, *speechiness* y *tempo* no muestran relación clara con la popularidad.

MULTICOLINEALIDAD

Se detecta fuerte multicolinealidad entre *energy* y *loudness* (r = 0.76) y entre *energy* y *acousticness* (r = -0.71). Esto puede derivar en problemas en la interpretación de nuestros modelos

MISSINGS

a) **Estrategia Implementada: MICE**
Se seleccionó el método MICE (Imputación Multivariada).
Por qué: Es una estrategia robusta que maneja eficazmente la diversidad de variables (numéricas, categóricas) y preserva las relaciones e interacciones entre ellas.

b) **Resultados Relevantes**
Se imputaron un total de 51,428 valores faltantes.
Esto resultó en un conjunto de datos completo, listo para el modelado.

c) **Decisión**
Se valida el uso de MICE como la estrategia óptima para mantener la integridad estructural de los datos sin introducir sesgos.

TRATAMIENTO DE OUTLIERS

a) **Estrategia Implementada: Isolation Forest**
Se utilizó el algoritmo Isolation Forest para la detección de anomalías.
Tras la detección, se realizó una inspección manual de los casos señalados.

b) **Resultados Relevantes**
El algoritmo identificó 14 outliers. La inspección confirmó que no eran errores de entrada, sino casos atípicos genuinos: canciones con combinaciones de características muy poco comunes (ej. instrumentales silenciosas, grabaciones en directo de baja calidad).

c) **Decisión: Añadir pesos**
Se decidió añadir pesos los 14 registros identificados. Aunque eran valores "reales", queremos que tengan menos influencia en los modelos que el resto de observaciones.

Objetivo: Evitar que el modelo se sesgue aprendiendo patrones de casos extremos y, así, crear un modelo final más robusto y generalista.

INGENIERÍA DE VARIABLES

Estrategia y decisiones

Transformación (Manejo de Asimetría)

Decisión: Variables con alta asimetría positiva o exceso de ceros distorsionan los modelos.

Estrategia: Se crearon nuevas variables _log aplicando la transformación Logarítmica (log1p) para normalizar su distribución.

Creación de Features

es_instrumental: Variable binaria (0/1) para capturar la señal de "Sí/No" de las canciones instrumentales.

valence_binned: Variable categórica (baja/media/alta) creada por discretización para encontrar puntos de corte claros en audio_valence.

Encoding y Escalado

Encoding: Se generaron variables Dummies (OHE) para key, time_signature y la nueva valence_binned.

Escalado: Se crearon nuevas variables _scaled aplicando Estandarización a todas las numéricas (danceability, loudness, y todas las _log).

INGENIERÍA DE VARIABLES

Base de datos preprocesada

Variable Original	Problema Detectado	Nuevas Features Creadas (Ejemplos)
instrumentalness	Asimetría / Exceso de Ceros	instrumentalness_log (Continua) es_instrumental (Binaria 0/1)
key, time_signature	Categórica Nominal	key_C#, time_sig_4, ... (Dummies)
audio_valence	Distribución Plana	valence_binned_media, valence_binned_alta (Dummies)
song_duration_ms	Asimetría Positiva	song_duration_ms_log (Transformada)
danceability	Escala (0-1)	danceability_scaled (Z-score)
loudness	Escala (dB neg.)	loudness_scaled (Z-score)
..._log (Todas)	Escala post-transform.	..._log_scaled (Z-score)

REGLAS DE ASOCIACIÓN

Workflow: Enfoque Basket Analysis

El objetivo fue transformar el problema de regresión en un análisis de cesta de compra para descubrir qué *combinación* de características (items) conduce a un nivel de popularidad.

1. Discretización: Todas las variables numéricas se convirtieron en items con 3 niveles ("Low", "Medium", "High").

Variables	**Método de Discretización**	**Justificación**
liveness, acousticness, speechiness, song_duration_ms	**'frequency'** (Cuantiles)	Datos con alta asimetría (skewed). Asegura que cada *bin* (item) tenga el mismo número de canciones.
loudness, danceability, energy, tempo, audio_valence, **song_popularity**	**'intervals'** (Anchura Fija)	Datos más simétricos. Crea *bins* de igual tamaño (ej. Pop 0-33, 34-66, 67-100).

2. Objetivo Predictivo: Buscamos reglas donde el Consecuente (RHS) sea un nivel de popularidad, con una buena confianza.

REGLAS DE ASOCIACIÓN

Resultados

Selección de Umbrales:

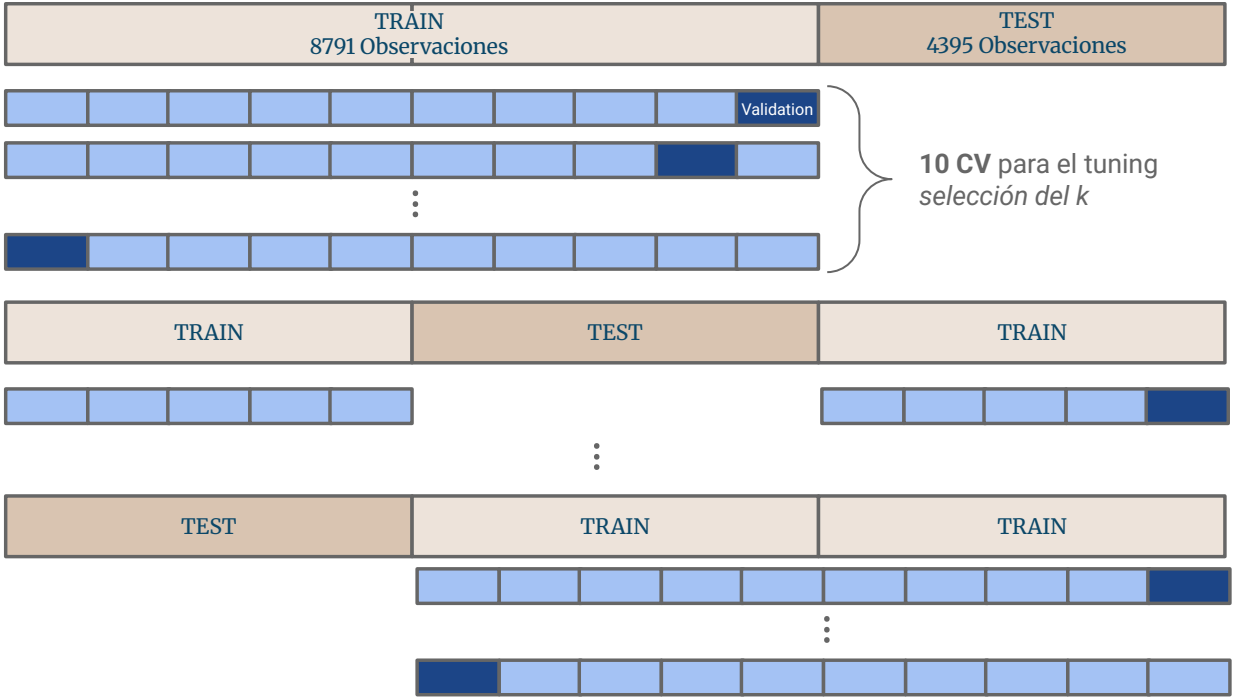
- **min_support** = 0.1: Establecido tras analizar la frecuencia de items.
- **min_confidence** = 0.4: Se fijó un umbral bajo intencionadamente.
- Hallazgo: Con min_confidence > 0.4, el algoritmo no generó NINGUNA regla que predijera song_popularity = High.

Conclusión: Reglas Débiles: Las reglas encontradas para Pop_Media o Pop_Baja presentaron valores de **Confianza y/o Lift demasiado bajos** para ser estadísticamente robustas.

Decisión:
Los hallazgos actuales (débiles) no justifican un cambio en la metodología de regresión definida

El análisis de reglas de asociación fue un paso de exploración válido, pero confirma que un **enfoque de regresión (k-NN)**, que *sí* puede manejar las sutilezas de las variables continuas, es el camino correcto para este problema.

PREPARACIÓN PARA EL MODELADO



→ 3x10-Cross Validation para obtener una estimación robusta del RMSE

→ Se probarán modelos con diversas combinaciones de variables originales y variables creadas o transformadas durante la ingeniería de características

PRIMEROS MODELOS – KNN

- **Model 1 w/o weighted distances**: utilizando la función VIM::kNN() (diseñada originalmente para imputación mediante kNN).

- **Model 2 with weighted distances**: aplicando FAMD como paso de preprocesamiento y entrenando posteriormente con caret::train(method = “kknn”).

Modelo	Estrategia de k-NN	RMSE (Error)
Modelo 1	Estándar (no ponderado) *(VIM::kNN)*	22.00
Modelo 2	**Distancias Ponderadas** *(FAMD + caret::kknn)*	**21.45**

En el Modelo 1, el proceso del tuning sugirió un **valor óptimo de k ≈ 100**, obteniendo un **RMSE de 22.00**, en el Modelo 2 un **k ≈ 110** con un **RMSE de 21.45**.
Por lo tanto, el Modelo 2 presenta un **rendimiento ligeramente superior**.

La transformación logarítmica de las variables con alta asimetría positiva no mostró mejoras en el desempeño del modelo.

CONCLUSIONES Y SIGUIENTES PASOS

Conclusiones

- Las reglas de asociación no han mostrado ninguna relación lo suficientemente razonable como para conseguir reducir la dimensionalidad.
- El modelo de regresión K-nn, con un valor óptimo cercano a 100 y dando pesos a las observaciones, ha conseguido modelar las variables con RMSE = 21.44
- En el K-nn las transformaciones para mejorar asimetrías en las variables no han resultado en una mejora del RMSE del modelo.

Siguientes pasos

- Utilizar otros modelos combinando variables originales y transformadas para conseguir el mejor modelo, el cual se optimizará y comparará usando MAPE (Objetivo de la competición).
- Probar Modelos Avanzados (Árboles).
- Entender los coeficientes del modelo ganador para estudiar el comportamiento de las variables.