Instituto Tecnológico y de Estudios Superiores de Monterrey

TC3006C: Inteligencia Artificial Avanzada para la Ciencia de Datos (Gpo 102)

ML-6: Proyecto ML - Limpieza y preparación

Integrantes:
A00835733 - Rolando Ruiz Martínez
A01285367 - Natalia Olvera Ortiz
A01425341 - Carol Jatziry Rendon Guerrero
A01571463 - Ericka Sofía Rodríguez Sánchez
A00835576 - Marcos Renato Aquino Garcia

Profesor:
Alder López Cerda

Monterrey, N.L. a 23 de agosto de 2025

La fase de limpieza y transformación de datos constituye un paso esencial dentro del proceso de análisis, ya que permite garantizar la calidad y consistencia de los datos antes de aplicar algún modelo estadístico o de *Machine Learning*. En este entregable se cubrirán algunas etapas clave como: la carga y exploración inicial de los datos para estudiar su estructura y posibles problemas; el tratamiento de valores faltantes, donde se busca reducir el sesgo por información incompleta; el manejo de *outliers*, evaluando si son errores o comportamientos relevantes en las canciones; la codificación de las variables categóricas, necesaria para que sea posible procesarlas con algoritmos; por último, el escalado y/o normalización, que tiene como objetivo homogeneizar las magnitudes de variables numéricas y que los modelos tengan un desempeño sin desbalances.

1. Carga y exploración inicial

El proceso comenzó importando las librerías necesarias para cargar el archivo de tipo CSV en un *DataFrame* de *Pandas*. Luego se realizó una exploración inicial de los datos para diferenciar las columnas numéricas de las categóricas, de la cual se obtuvo el resultado de 15 columnas numéricas y 6 categóricas.

Para complementar la información del set original, se desarrolló un código para establecer una conexión con la API de *Spotify for Developers*. Usando el identificador de cada canción, se consultó la fecha de lanzamiento, y para contar con un formato estándar, se conservó solamente el año en una columna de tipo numérico llamada *'release_year'*.

Esta primera sección es un paso crucial, ya que además de complementar el set con nueva información, nos permite comprender su estructura antes de comenzar a manipular los datos. Al resaltar el tipo de dato que está contenido en cada columna, se puede elegir de manera informada qué técnica de limpieza y preprocesamiento es la indicada.

2. Tratamiento de valores faltantes

En el dataset fueron encontrados datos faltantes en distintas columnas, como lo mostrado en el Cuadro 1.

Cuadro 1: Conteo de valores nulos encontrados por columna.

Columna	**Datos nulos**
artists	1
album_name	1
track_name	1
release_year	252

Para el valor nulo encontrado en *'artists'*, *'album_name'* y *'track_name'*, se identificó que proviene de un mismo registro. No obstante, se trata solo de información identificadora para

la canción, por lo que se mantuvo el registro y se eliminaron dichas columnas. En cambio, para *'release_year'*, se eliminaron los registros que contuvieran el valor nulo, puesto que representa un mínimo porcentaje en la base de datos.

3. Tratamiento de valores duplicados

Durante el proceso de limpieza del dataset se identificaron un total de 450 registros duplicados, los cuales fueron eliminados para garantizar la consistencia de la información. Después de aplicar este tratamiento, el conjunto de datos quedó conformado por 113,131 registros únicos, lo que permite trabajar con una base de datos depurada y libre de repeticiones.

4. Tratamiento de *Outliers*

En el tratamiento de *outliers* es importante realizar una visualización de las distribuciones de valores numéricos antes de realizar cualquier transformación. Debajo en la Figura 1 se muestra una visualización general de las variables numéricas no categóricas de la base de datos.

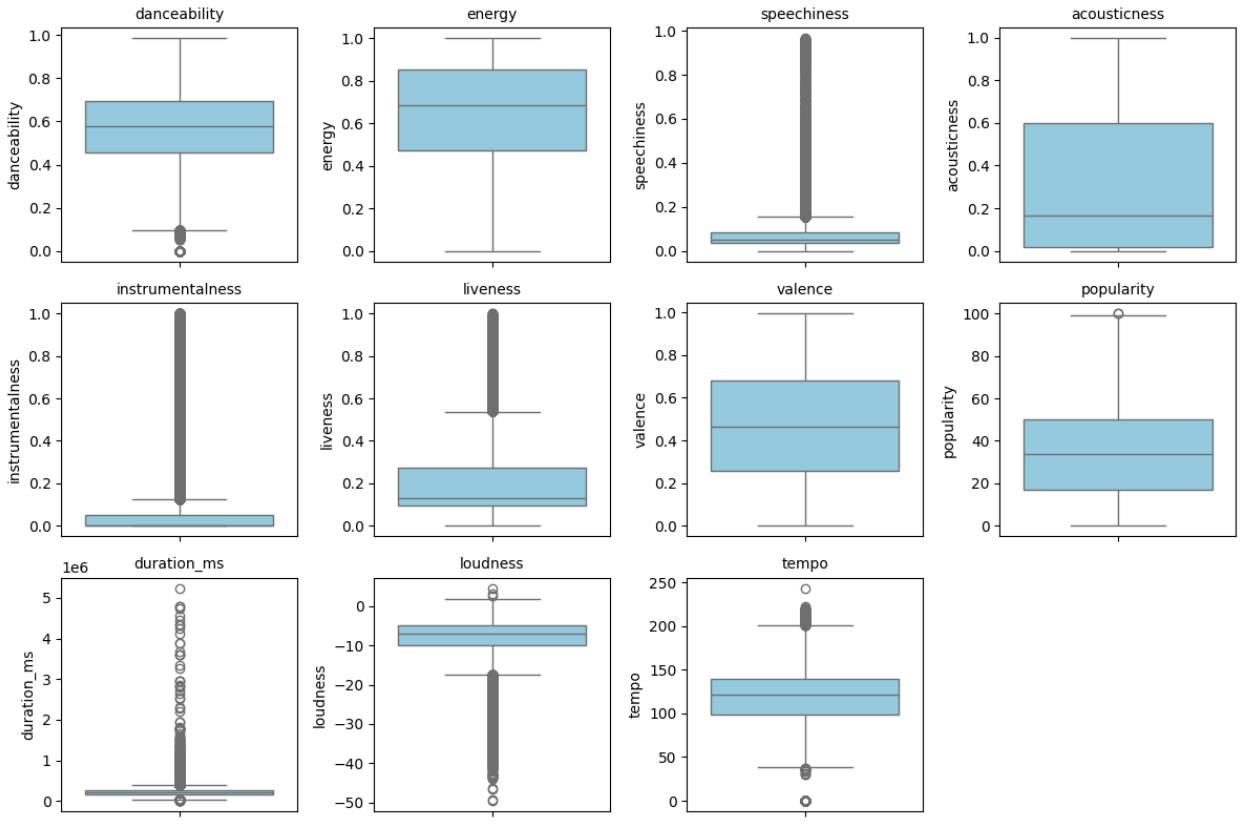

Figura 1: Diagramas de caja y bigote de las variables numéricas

Las variables donde resaltan los valores atípicos son *'speechiness'*, *'instrumentalness'*, *'liveness'*, *'loudness'* y *'duration_ms'*. Para las primeras cuatro variables, se consideró ignorarlas debido a que son datos que caracterizan cierto género y representan la riqueza y variedad en la música, como lo es el género clásico con *'instrumentalness'* o rap con *'speechiness'*. No obstante, con *'duration_ms'* es un caso distinto, puesto que existen registros de canciones que superan los 20 minutos, incluso la hora. Por tanto, se optó por mantener una duración que sea representativa en el consumo general, eliminando aquellos que duren menos de 30 segundos (interludios, intros) y mayores a 15 minutos (experimentales y grabaciones completas de conciertos).

Al eliminar los registros que salgan de los límites acordados de duración, se eliminaron 169 filas adicionales. En la Figura 2 se muestra cómo cambia la distribución de la variable con este acotamiento de la duración.

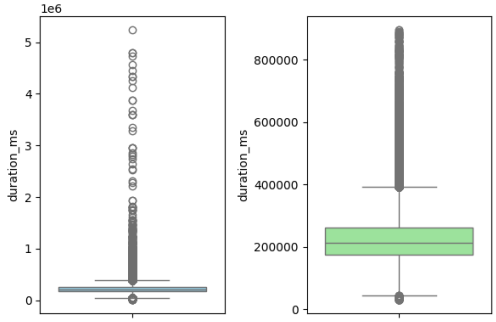

Figura 2: Comparación de variable *'duration_ms'* antes y después de imponer los límites de duración.

5. Codificación de Variables

De igual manera, para la codificación de variables categóricas, primero se revisaron las distribuciones de ellas para evaluar si es necesaria la modificación de alguna de ellas. Sus visualizaciones son encontradas en la Figura 3.

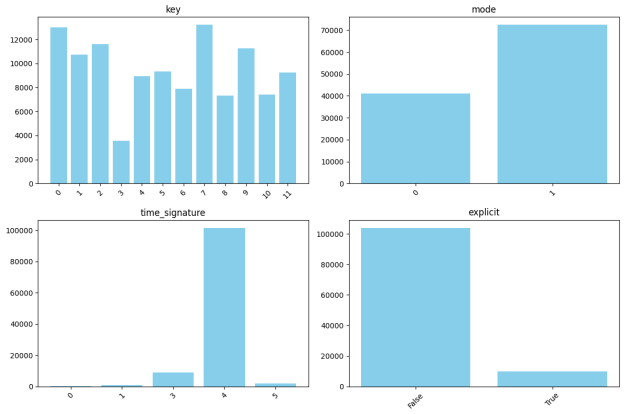

Figura 3: Histogramas de variables categóricas

Como se puede observar en la Figura 3, la variable *'mode'* ya está en formato binario, lo cual nos ayuda, y la variable *'explicit'* que estaba en formato booleano, también la codificamos en binario para mayor facilidad de manejo. Así también, se realizó un *One Hot Encoding* en la variable *'key'* debido a que es nominal y tiene un balanceo significativamente similar en cada valor.

Por otro lado, se decidió modificar la variable '*time_signature*' debido a que para futuros análisis, será de mayor valor reducir la cardinalidad a un estado más balanceado. Por esto último, se deriva la variable *'time_signature_4'* donde 1 indica que *'time_signature'* es 4 y 0 que no.

También, se encuentra en la Figura 4 el histograma que representa la distribución de la variable de *'release_year'*.

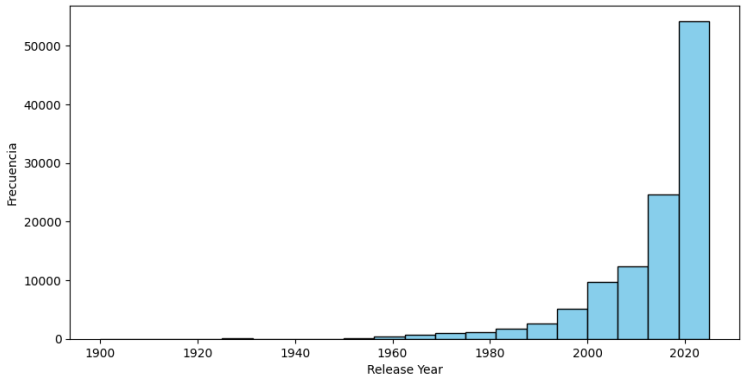

Figura 4: Histograma release_year

Arriba en la Figura 4 se puede apreciar cómo la cantidad de canciones por año va aumentando exponencialmente, es por esto, y por la facilidad de manejo, que se optó por dividir el *'release_year'* en una variable binaria, siendo de 2015 y años anteriores valor de 1, y de 2016 en adelante 0. Esto nos permitirá hacer análisis de canciones en distintas épocas, con un set de datos más balanceado.

Debajo en la Figura 5 se muestran las variables derivadas del análisis en su nueva distribución.

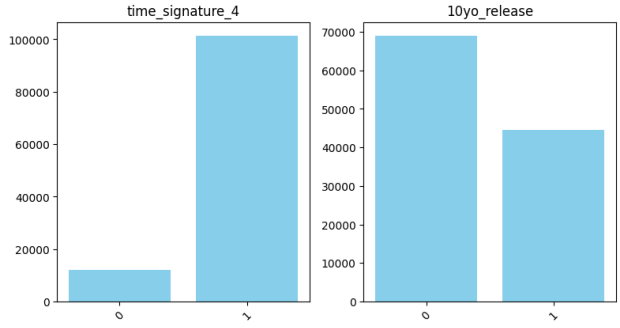

Figura 5: Histograma del balance de las variables

En la Figura 5, primero se observa el balance de *'time_signature_4'* donde está muy sesgado hacia el 1 que significa que sí es 4. A pesar de estar desbalanceado, es importante ver la diferencia con la reducción de variables implementada, en donde antes existían 5 valores posibles, y ahora es uno binario lo cual facilita la interpretación y utilización de la variable. Por el otro lado, tenemos el histograma de *'10yo_release'* en donde al dividirlo como mencionado antes, se crea un mayor balance dentro de la variable sin tener año por año de cada canción.

6. Escalado/Normalización

Se seleccionaron las columnas de *'popularity', 'duration_ms' 'loudness', 'release_year'*, y *'tempo'* para realizar un escalado. Se utilizó el método de *StandardScaler* de la librería *scikit-learn*, este método estandariza los datos para que tengan una media de cero y una desviación estándar de 1. Posteriormente, se realizó una verificación de las estadísticas descriptivas, las cuales demostraron un escalado exitoso.

Se eligieron las variables antes mencionadas para asegurarse de que, debido a la escala, ninguna domine sobre la otra y que no genere un sesgo. Este paso es importante en especial para aquellos modelos que son sensibles a la escala de las características, tales como la regresión logística. Las demás columnas numéricas, como *'daceability', 'energy', 'speechiness', 'acousticness', 'instrumentalness', 'liveness'* y *'valence'*, ya se encontraban en una escala del cero al uno. Por lo tanto, no requerían de un escalado.

7. Hallazgos y conclusiones

Durante esta etapa de limpieza y preparación de los datos, se logró dejar un dataset listo y limpio para ser usado en fases posteriores. De primero, se analizaron los nulos que se pueden ver en la tabla 1, en donde únicamente las columnas de *Artists, album name, track name y release year* tenían valores nulos, los cuales, debido a su pequeño porcentaje, se decidieron eliminar.

Luego, se analizaron los outliers, en donde vimos que bastantes variables tenían outliers, pero se decidieron tratar de distintas maneras. Las variables de *speechiness, instrumentalness, liveness y loudness*, se decidió ignorar, debido a que son datos que caracterizan las canciones y no pueden ser ignorados. Por otro lado, la variable de *duration ms* tuvo que ser tratada distinta, ya que existían ejemplares que duraban mucho tiempo o muy poco, lo cual se sale del rango de una canción, por lo que se decidió eliminar aquellos menores a 30 segundos y mayores a 15 minutos.

Durante la codificación de variables categóricas, decidimos pasar las variables de *time signature y explicit* a binarios y a *key* se le aplicó one hot encoding, ambos casos para tener un mayor balance y usabilidad de estas variables categóricas.

Para el escalado, se escogieron las variables de *popularity, duration ms, loudness, release year y tempo* utilizando el método de standard scaler de Scikit learn, el cual hace que los datos tengan media de 0 y desviación estándar de 1.

En resumen, la limpieza y preparación de datos fue hecha de acuerdo con los lineamientos necesarios, y estos están listos para los siguientes pasos. Es ideal mencionar que con estos datos, se espera que no existan sesgos tan grandes, ni datos que afecten futuros modelos a realizar.