Informe de Calidad de Datos

Lohanna Aguirre

8 de enero de 2024

1. Introduction

Los datos rara vez son perfectos; de hecho, la mayoría contiene errores de codificación, valores perdidos y otras inconsistencias que pueden complicar el análisis. Para evitar problemas al generar estos análisis, se recomienda realizar un preprocesamiento en el cual se valide la calidad de los datos. Este paso puede abordar diversas dimensiones, como la completitud de los datos, la consistencia entre variables, la validez de los datos, la precisión de los valores, la unicidad de los registros y la integridad referencial, entre otros aspectos.

En este documento, nos enfocaremos en los resultados del análisis de calidad de datos para el conjunto de datos denominado `'taylor_swift_spotify.json'`.

2. Hallazgos

Teniendo en cuenta las seis dimensiones del análisis de calidad de datos, se examinó cada una de ellas y se les aplicaron las técnicas que mejor se adaptaban a los datos, considerando la información disponible. Además de este análisis, se llevó a cabo una exploración de datos que también contribuyó a identificar ciertas anomalías.

2.1. Anomalías

- A pesar de que el campo `'track_popularity'` es un número entre 0 y 100, se han encontrado valores máximos y mínimos de 152 y -92, respectivamente.
- Algo similar sucede con el campo `'artist_popularity'` es un número entre 0 y 100, y este cuenta con un uncio valor de 120.
- El campo `'album_release_date'` hay una fecha que aun no ha sucedido como lo es '2027-05-26' y una fecha que es muy antigua teniendo en cuenta al artista '1989-10-24'.

2.2. Integridad

El conjunto de datos consta de 27 columnas y 538 filas. En este contexto, se identifican 12 columnas que contienen valores nulos.De las 27 columnas, 6 presentan un único valor nulo, mientras que 5 contienen valores que oscilan entre 2 y 8. Por último, existe una columna con 62 registros nulos.

Esto implica que el 44,44 % de las columnas incluyen registros nulos, los cuales representan menos del 1 % en relación con la cantidad total de datos.

Columna	Cantidad de Nulos	Porcentaje de nulos
track_id	8	1,5
track_name	7	1,3
audio_features.danceability	2	0,4
audio_features.energy	2	0,4
audio_features.key	1	0,2
audio_features.loudness	2	0,4
audio_features.speechiness	1	0,2
audio_features.acousticness	1	0,2
audio_features.liveness	1	0,2
audio_features.tempo	1	0,2
audio_features.time_signature	1	0,2
album_name	62	11,5

Cuadro 1: Nulos

2.3. Unicidad

Al validar el criterio de unicidad en nuestro conjunto de datos, columna por columna, encontramos algunos valores nulos. Sin embargo, al analizar cada fila con respecto a la información de todas las columnas, identificamos 18 registros duplicados. Esto significa que el 96,6 % de los datos son únicos.

$$\frac{538 - 18}{538} * 100 = 96{,}65\,\%$$

2.4. Validez

De las 27 columnas del dataframe, se descubrió que cinco de ellas no eran detectadas correctamente por Python al ser leídas, de acuerdo con la documentación del API.

Esto implica que el 18 % de los datos no coinciden en su tipo de dato.

3. Resultados

Luego de llevar a cabo un exhaustivo análisis de calidad de datos, se implementaron diversos procesos y técnicas con el fin de evaluar la integridad y consistencia de nuestro conjunto de datos. Durante este proceso, se emplearon métodos de validación específicos para identificar valores atípicos.

Es importante señalar que la elección de dimensiones para la validación no implica que todas deban aplicarse uniformemente a todos los datos. La selección de la dimensión a evaluar depende de la estructura particular de nuestro conjunto de datos. Además, cada una de estas dimensiones debe ser validada en función de su relevancia para el conjunto total de datos, reconociendo y priorizando la importancia relativa de cada aspecto en el análisis global.