Informe parcial

2022-06-10

Table of contents

CARÁTULA

- **Curso:** Análisis de Datos
- **Profesor:** Stefany Neciosup
- **Código del curso:** 1INF03
- **Fecha de Entrega:** 11/06/2022
- **Integrantes:**
  - Richle Gianotti, Renzo Ernesto - 20180368

- Cornejo Ramírez, Lucio Enrique - 20192058
- Vivas Alejandro, Claudia Mirela - 20141150
- Mejia Padilla, Andrea Adela - 20180824

INTRODUCCIÓN

Para entender la dinámica de la industria musical, antes de nada, es necesario saber que no se trata de una sola, sino de varias, diferentes, estrechamente relacionadas entre sí, pero que parten de lógicas y estructuras distintas. La industria musical en su conjunto vive de la creación y la explotación de la propiedad intelectual musical. Compositores y letristas crean canciones, letras y arreglos que se interpretan en directo sobre el escenario, se graban y distribuyen a los consumidores. Esta estructura básica ha dado lugar a tres industrias musicales centrales: la discográfica, centrada en la grabación de música y su distribución a los consumidores; la de las licencias musicales, que sobre todo concede licencias a empresas para la explotación de composiciones y arreglos, y la música en vivo, centrada en producir y promocionar espectáculos en directo, como conciertos, giras, etcétera.

En la actualidad se vienen realizando muchos estudios aplicando el campo de Data Analytics para poder obtener resultados en muchas industrias, y la musical no es ajena a esta moda. Es por esto que decidimos desarrollar el presente trabajo para así descubrir las características fuertemente asociadas a las canciones más populares en la aplicación de música Spotify y poder usar estos rasgos, que están conectados con los estados de ánimo, fechas de lanzamiento, nombre de las canciones, entre otros; para analizar el por qué de esta popularidad y poder simular que una empresa de la industria musical quiera emplear un modelo analítico que le permita saber el índice de popularidad de una canción antes de que esta sea soltada al mercado.

Entre los puntos que queremos saber:

- ¿Cuáles son los tracks más populares en Spotify?
- ¿Qué características en común tienen las canciones más populares de Spotify?
- ¿Existe una correlación entre la popularidad y alguna característica de las canciones?
- ¿Cuánto debe durar un track según los estándares de la actualidad?
- ¿Cuál es la correlación entre diferentes tracks que son populares?

Dataset

El dataset elegido se obtuvo utilizando el servicio API de Spotify, el cual nos permite obtener la información de las canciones existentes en tal plataforma. El conjunto de datos que estamos empleando consta inicialmente de 20 columnas, las cuales representan variables como el nombre de la canción, su popularidad, duración, el artista, la fecha de lanzamiento, bailabilidad de la canción, su energía, volumen sonoro, etc.

CAPÍTULO I: COMPRENSIÓN DEL NEGOCIO / ESTUDIO

Descripción del problema

En la actualidad se sabe qué canciones son las más populares en general observando la cantidad de reproducciones que tienen cada canción, sin embargo la industria de la música es un negocio y como tal una empresa discográfica siempre busca que su canción sea popular para que esta genere ingresos a la empresa. Si pudiéramos saber que características hacen popular a una canción entonces se podrían usar a favor para generar canciones que sean del agrado del público oyente y como consecuencia pueda otorgar un mayor beneficio económico.

Objetivo del negocio/estudio

Nuestro objetivo principal es el de indagar, descubrir y utilizar las características que definen la popularidad de las canciones presentes en nuestro dataset.

Variable objetivo para el negocio

- Las variables que consideramos podrían ser más relevantes para el negocio son:
  - popularity (Popularidad)
  - duration_ms (La duración del track en milisegundos)
  - danceability (La capacidad de baile del track)

CAPÍTULO II: FUENTES DE INFORMACIÓN

- Origen de los datos del proyecto
- Descripción del universo y muestra (incluyendo descripción en espacio y tiempo)
- Descripción y entendimiento de variables (medidas de resumen y gráfico, según tipo)

CAPÍTULO III: PREPROCESAMIENTO DE DATOS

En ese capitulo se detalla el proceso de preprocesamiento de los datos. Por otro lado, la variable popularity es la variable dependiente del estudio por lo cual esta variable no debe pasar por ningún tratamiento de outliers, vacíos y/o transformación

Selección de registros y atributos

Excluimos la variables “artist”, pues esta no aporta información para la predicción de popularidad de canciones.

Tratamiento de datos atípicos

Si bien se ahorraría tiempo realizar el análisis de valores outliers de las variables numéricas de forma automática donde se usaría el límite superior e inferior teórico, para poder realizar un análisis minucioso, las variables numéricas se analizarán de forma particular. El proceso que seguimos fue cambiar el coeficiente que acompaña al rango intercuartil al momento de definir los limites superiores e inferiores.

Loudness

Describe la sonoridad general de una pista en decibeles. Los valores de sonoridad se promedian en toda la pista y son útiles para comparar la sonoridad relativa de las pistas. La sonoridad es la cualidad de un sonido que es el principal correlato psicológico de la fuerza física (amplitud). Los valores suelen oscilar entre -60 y 0 db. Se observa que la distribución de la variable se sesga hacia el lado derecho, en este caso usamos a 1.7 como coeficiente. El porcentaje de valores atípicos es 9% del total de los datos de esta columna. Por lo que es posible generar imputaciones.

Instrumentalness

Predice si una pista no contiene voces. Los sonidos "Ooh" y "aah" se consideran instrumentales en este contexto. Las pistas de rap o de palabras habladas son claramente "vocales". Cuanto más se acerque el valor de instrumentalización a 1,0, mayor será la probabilidad de que la pista no tenga contenido vocal. Los valores superiores a 0,5 representan pistas instrumentales, pero la confianza es mayor a medida que el valor se acerca a 1,0. Se observa que la mayoría de los datos están concentrado alrededor de cero, es decir la distribución esta sesgada a la izquierda, en este caso usamos a 3 como coeficiente. El porcentaje de valores atípicos es mayor a 20% del total de los datos de esta columna. Por lo que es no posible generar imputaciones.

Liveness

Detecta la presencia de una audiencia en la grabación. Valores más altos de liveness representan una probabilidad incrementada de que la pista haya sido realizada en vivo. Un valor de 0.8 provee una probabilidad fuerte de que la pista sea en vivo. Se observa que los datos están concentrados alrededor de 0.1 aproximadamente; no obstante tales datos muestran un fuerte sesgo hacia la derecha, e incluso presentan un ligerada bimodalidad. Usamos un coeficiente de 1.3 para determinar a los valores outliers. El porcentaje de valores atípicos es de 9.51%.

Valence

Es medida desde 0.0 hasta 1.0 que describe la positividad musical expresada por la pista. Las pistas con un sonido de alta valence suenan más positivos (e.g., felices, alegres, eufóricos), mientras que las pistas con una baja valence suenan más negativas (e.g., tristes, deprimentes, enojadas). Se observa que al distribución se asemeja a uns distribución unirfore , donde parace que cada valor tomado por esta variable se acerca a un mismo nivel de densidad. Usamos un

coeficiente de 1.4 para determinar a los valores outliers y el porcentaje de valores atípicos es de 9%.

Tempo

Es el tempo general estimado de una pista en beats por minutos (BPM, por sus siglas en inglés). En terminología musical, el tempo es la velocidad o ritmo de una pieza dada y se deriva directamente de una duración promedio de beat. Se observa que los datos presentan múltiples modas, aproximadamente alrededor de los valores 75, 100, 140, 175. Estos datos presentan un ligero sesgo hacia la derecha. El porcentaje de valores atípicos para esta variable es 8.90%.

Duration_ms

Es la duración de una pieza en milisegundos. Respecto a la distribución de esta variable, se observa que la distribución de los datos es en principio aparentemente degenerada (los datos están virtual o totalmente reunidos en un punto). No obstante, tal aparente degeneración se explica por el hecho de que existe una serie de piezas cuya duración puede ser extremadamente larga. En este caso, si bien los valores atípicos representan solo el 8.99% del total de la muestra de valores para esta variable, tal cantidad de valores atípicos, dados sus valores extremadamente altos, tienen un efecto visual altamente significativo sobre la interpretabilidad de la distribución de los datos.

Danceability

La danceability describe qué tan adecuada es una pista para bailar, basada en una combinación de elementos musicales incluyendo el tempo, la estabilidad rítmica (rhythm stability), la fuerza del beat (beat strength), y una regularidad general. Un valor de 0.0 es menos danceable 1.0 es máximamente danceable. Se observa que los datos se concentran alrededor de 0.65, y presentan un ligero sesgo hacia la izquierda. Los valores atípicos representan 8.216% de la muestra,para calcularlos se usó el coeficiente 1.6 que acompaña al rango intercuartil.

Energy

La energía es una medida desde 0.0 a 1.0 y representa una medida perceptiva de intensidad y actividad. Típicamente, las pistas energéticas se sienten rápidas, de volumen alto (loud), y ruidosas. Por ejemplo, el death metal tiene una alta energía, mientras que el preludio de Bach da un puntaje bajo en la escala. Características perceptivas contribuyen a este atributo incluyen rango dinámico, el volumen percibió, el timbre, el onset rate, y la entropía general. Se observa que los datos son ligeramente multimodales. con los datos ligeramente más concentrados alrededor de 0.35, 0.55, 7.00 y 0.85. Los valores atípicos para esta variable representan 9.32% de la muestar y para calcularlos se usó el coeficiente 1.4 que acompaña al rango intercuartil.

Speechiness

Esta variableetecta la presencia de palabras habladas en una pieza. Mientras más contenido hablado presenta una grabación (e.g., talk show, audio book, poesía), más cercano a 1.0 será el valor del atributo. Los valores superiores a 0.66 describen piezas que probablemente estén

hechas enteramente de palabras habladas. Valores entre 0.33 y 0.66 describen piezas que podrías contener tanto música como una parte oral, ya sea en secciones o en capas, incluyendo casos como el rap. Valores menores a 0.33 con mayor probabilidad representan múscia y otras piezas non-speech-like. De manera similar al caso de la variable Duration_ms, aunque en medida mucho menor, esta variable, Speechiness, presenta en principio una aparente distribución degenerada, concentrada alrededor 0. No obstante, tal aparente naturaleza se debe al 10.39% de valores atípicos que presenta la muestra de datos para esta variable, cuyos niveles están concentrados alrededor de 1. En otra palabras, aproximadamente el 88% de los datos se concentra alrededor de 0.0, mientras que el 10.39% lo hace alrededor de 1.0.

Acousticness

Una medida de confianza desde 0.0 hasta 1.0 sobre si la pieza es acústica. 1.0 representa alta confianza en que la pieza es acústica (>=0 | <= 1). Se observa que los datos se concentran en su mayoría alrededor de ambos valores extremos. La distribución de los valores para esta variable es en tal sentido bimodal. Los valores atípicos representan un 8.76% del total de valores de la muestra para esta variable.

Tratamiento de datos vacíos

Los datos que disponemos no tienen valores vacíos, por lo que obviamos este análisis.

Creación y transformación de variables

Inicialmente la base de datos con la que contábamos tenía una reducida cantidad de columnas, de las cuales algunas no aportaban con información relevante para el objetivo del negocio. En ese sentido, transformamos las variables name, time_signature y realice_data. De las cuales obtenemos información que podrán servir para la predicción.

Name

Contiene el nombre de la canción, a partir de esta se generan dos variables:

- Name_lenght : contabiliza al cantidad de caracteres string del nombre de la canción omitiendo los espacios vacíos entre palabra y palabra.
- Words_name: contabiliza la cantidad de palabras que están presentes en el nombre de una canción.

Es importante recordar, que luego de generar ambas variables, se tiene que eliminar la variable "name" para no caer en el problema de multicolinealidad.

Realise date

Indica la fecha del lanzamiento de la canción incluyendo el año, mes y día. Apartir de esat variable se crearon cuatro variables:

- Release_year: año en el que se publicó la canción
- Release_month: mes en el que se publicó la canción
- Release_days: día en el que se publicó la canción
- Release_trim: trimestre en el que se publicó la canción

Posteriormente, eliminamos la variable realise date.

Time_signature

Esta variable contiene información sobre el compás, este variable toma valores de 3 a 7, asi discretizamos esta variable para que tome el valor de 0 si los valores son mayores iguales a 0 y menores que 4; y tome el valor de 1si los valores toman valores mayores iguales a 4.

Descripción de variables listas para el modelamiento

Análisis de correlación

Se observa que los colores que adoptan las casillas de la matriz de correlación son rosados, lo cual indica que las variables tienen un coeficiente de correlación que se encuentra entre 0.2 y 0.4, es decir el nivel de correlación entre variables numéricas predictoras no es muy alta.