**Abstract**

El dataset "Spotify\_Track\_Dataset” está compuesta por una muestra de 114.000 canciones que fueron seleccionadas del servicio de streamming de música Spotify, el cual posee más de 80 millones de canciones y podcasts. Este dataset posee 21 columnas que describen cada canción, entre dichas columnas se encuentra el nombre de la canción, su interprete, álbum, duración y una serie de parámetros numéricos que dan información sobre el volumen, la energía que transmite, el carácter instrumental, la presencia de palabra hablada o cantanda, entre otras.

Por lo visto en las dos entregas anteriores y que reforzaremos en la actual entrega, el dataset posee potencial para establecer un sistema de clasificación de las canciones que no se base solamente en su género musical, sino que incluya criterios tales como música bailable, para relajación, apta para menores, etcétera. Tal sistema de clasificación da pie para luego crear un sistema de recomendación que ofrezca una lista de canciones que se adecue a los gustos y preferencias particulares de cada usuario de la plataforma Spotify.

A los fines de evaluar la factibilidad de realizar tal sistema de clasificación se plantean una serie de preguntas de interés que buscarán ser respondidas en la presente entrega.

* ¿Es la muestra representativa del total de las canciones del servicio Spotify?

Dado que no se dispone de información de cómo se seleccionaron las canciones que están presentes en el dataset, resulta relevante realizar un análisis exploratorio buscando identificar como fueron elegidas las canciones. Esto es crítico ya que dicha muestra puede no ser una buena representación de la población, dando lugar a posibles sesgos.

* ¿Los datos están completos y bien formateados?

Si bien esta pregunta siempre estuvo en mente desde la primera entrega, en esta ocasión se hará un análisis un tanto más profundo, identificando la presencia de valores NaN o Null y chequeando que los datos categóricos tengan uniformidad, es decir, que no existan diferencias tales como “led zeppelin”, “led\_zeppelin”, “Led zeppelin”, “Led Zeppelin”, etcétera.

* ¿Hay diferencias estadísticas significativas entre los diferentes parámetros que clasifican las canciones?

Comprobar si las canciones poseen diferencias palpables entre sus parámetros es fundamental ya que dichas diferencias serán la base sobre la cual el sistema de clasificación va a operar. Se especula que a mayores diferencias más robusto será el sistema de clasificación.

* ¿Cuáles son las correlaciones en los diferentes parámetros?

Explorar las correlaciones entre las diferentes variables es básico y fundamental de todo análisis exploratorio de datos, en este dataset las correlaciones ya sean positivas como negativas nos ayudaran a seleccionar valores de parámetros deseables que cumplan con las preferencias de los usuarios. Por ejemplo, si un usuario prefiere canciones enérgicas posiblemente busque canciones con altos valores de tempo y valencia.

El sólo hecho de buscar responder estas preguntas nos servirá de guía para realizar el análisis exploratorio de los datos, el procesamiento de éstos y pistas sobre cuales algoritmos de machine learning utilizar para el sistema de clasificación.

Referencias

https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

https://newsroom.spotify.com/company-info/