# Informe TPE Fundamentos de la Ciencia de Datos

Integrantes:

* Dimuro, Agustín Nicolás
* Grillo, Agustín German
* Padilla, Tomás Agustín

Introducción

Durante el transcurso del siguiente informe trabajaremos sobre un conjunto de datos proveniente de Amazon Music, el cual contiene observaciones sobre canciones de la década de los años 1970 al 1979. Realizaremos sobre este conjunto el filtrado de datos los cuales puedan ser erróneos para luego poder utilizarlos con el fin de buscar respuestas a ciertas hipótesis que puedan ser útiles para obtener información acerca del comportamiento de las canciones en esa época. Si los datos nos lo permiten, trataremos de crear un modelo de regresión lineal el cual para utilizarlo como una herramienta que nos permita entender los vínculos de diferentes variables independientes contra otra variable. A su vez, queremos tratar de generar un modelo el cuál nos ayude a predecir los valores de la variable elegida.

Las hipótesis que planteamos y que vamos a estar analizando durante el transcurso del siguiente informe son las siguientes.

* Las canciones que tienen valores más altos de positividad y alegría hoy en día son las más populares.
* Existe una relación entre las canciones que son bailables y las que son instrumentales.
* Las canciones compuestas con 4 pulsaciones por compas durante la década del 1970 actualmente son más populares que aquellas que fueron compuestas con un distinto número.
* Las canciones en la primera mitad de la década tienen diferencias significativas con respecto al estilo, contra las de la segunda mitad. Poniendo el foco en el análisis sobre sus valores de energía, el grado de que tan acústica es la canción, las pulsaciones por minuto y el volumen promedio de la canción en decibeles.
* Las canciones que son más habladas con respecto a las demás son menos instrumentales.
* Las canciones que fueron grabadas durante un concierto en vivo tienen valores de energía más altos que las demás.
* Las canciones que tienen mayor grado de instrumentalidad son las que más duración tienen comparando con las demás canciones de la década.

Materiales

Al momento de realizar la exploración de los datos que estaban dentro del dataset encontramos que estaban en una muy buena condición, es decir, que poseían una alta confiabilidad en lo que pretenden representar de la población. Esto se debe a que no logramos identificar valores erróneos o incorrectos ni señales que indiquen que puedan ser falsos. Tampoco observamos posibles outliers ni valores fuera de rango.

A pesar de lo mencionado, nos encontramos con la columna “Popularity” la cual, en una primera impresión nos daba a entender que representaba la popularidad que había tenido la canción durante la década. Pero al tratar de contrastar ese significado con lo observado en los datos nos dimos cuenta que era imposible que la variable mencionada represente la popularidad en la década de 1970 ya que, canciones que están situadas en los primeros puestos en los rankings de popularidad de canciones de esa década tenían valores mucho menores que canciones las cuales fueron mucho menos populares. Por esto decidimos buscar una nueva explicación para los datos almacenados en esta variable y nos encontramos en Kaggle con un dataset de Spotify el cual poseía una columna “Popularity” la cual los valores calculados en ella se mueven entre 0 y 100, al igual que en dataset nuestro, y que los valores generados para cada canción son calculados gracias a un algoritmo el cual toma en cuenta la cantidad total de reproducciones de la canción y que tan recientes son esas reproducciones. Dado que no encontramos cómo funciona el algoritmo que calcula la popularidad de las canciones en Amazon Music, asumimos que usa uno similar o igual a Spotify, por lo que la variable “Popularity” la analizaremos como la popularidad actual de las canciones.

Para facilitar el posterior análisis de los datos decidimos modificar los valores de algunas variables. Una de estas es la columna “Duration” la cual poseía los valores de la duración de la canción en minutos y segundos, pero decidimos modificar a que solo sea almacenado en segundos dado que puede generar problemas a la hora de interpretar los resultados el que esta columna posea dos unidades.

Otra de las columnas que modificamos fue “Loudness” ya que, luego de buscar en distintos sitios web especializados en música, información sobre cómo se calcula el volumen promedio de la canción en decibelios, siempre era mostrado como un valor entero y no con múltiples valores decimales. Por lo tanto, decidimos quitarle esos valores decimales a la columna y dejarlos como valores enteros.

De igual manera que con la columna “Loudness”, la columna “Tempo”, que mide la velocidad o el ritmo de la pista medido en pulsaciones por minuto, posee valores que están representados con números que poseen múltiples decimales. Al comparar los datos con los sitios web especializados en música, los valores que estaban almacenados en el dataset eran correctos, solo que no poseían la parte decimal. Por lo tanto, decidimos modificar esta columna para que solo almacene números enteros.

Métodos / Resultados

Conclusiones

Referencias