LA RECETA PARA UN HIT EN SPOTIFY

Genesis Ivonne Vega Gámez
Comisión 39960
29 de agosto de 2023

Tabla de Contenido

Abstracto

Con la creciente competencia en la industria musical por ocupar los primeros lugares de las listas más populares de música y en conjunto con el uso de la tecnología y plataformas de streaming y aplicaciones virales como Tik tok, el conocer los atributos de las canciones más populares puede ser una ventaja. La aplicación de la ciencia de datos con sobre esta clase de datos puede ayudar a la creación de una “receta” para el siguiente hit musical del año. Por lo que en el presente trabajo se analizaron las características de las canciones más populares de los 2000s a 2019 para crear un modelo de machine learning que ayude a la identificación del atributo de género musical de las canciones.

Contexto comercial y analítico

Las plataformas de streaming de música permiten que las personas tengan acceso a toda la música posible y a crear playlists basadas en diferentes criterios como género, popularidad, artistas, etc. Los anteriores atributos (listas de ranking y playlists) se han convertido en los productos por excelencia que ofrece la plataforma de streaming de música Spotify. Debido a esto, se tiene toda la información acerca de las canciones del momento y las más escuchadas por el mundo a nuestra disposición.

Esto conforma una gran base de datos accesible por la API de Spotify. Con la cual se ha podido recabar el top de canciones de los años 2000 al 2019 de acuerdo al catálogo de la plataforma. El estudio de las canciones más queridas por el público junto con sus características puede resultar en la determinación de qué es lo que hace que una canción sea popular, lo cual se vuelve valioso para las empresas productoras musicales e incluso artistas independientes.
Se proporciona un archivo CSV que contiene estadísticas de audio de las 2000 pistas principales en Spotify de 2000 a 2019. Los datos contienen alrededor de 18 columnas, cada una de las cuales describe la pista y sus cualidades.

Audiencia y limites

Este análisis intenta determinar las características que comparten las canciones más escuchadas y amadas por el público por lo que es dirigido para cualquier disquera o artista que busque crear un hit musical.

Los gustos musicales de las personas son muy variados entre países por lo que es difícil encasillar las canciones más populares debido a los diferentes datos y estadísticas que se pueden obtener de Spotify en diferentes países.

Preguntas y objetivos

El objetivo principal del estudio es realizar un análisis de los atributos de las canciones para identificar si existen tendencias y patrones dentro de estos para contestar a la pregunta: ¿qué atributos comparten las canciones más populares?

Las preguntas secundarias que se intentarán responder son:

- ¿Qué artistas lanzan canciones que se vuelven populares?
- ¿Cuáles son los géneros que más le gustan al público?
- ¿La duración de una canción importa en el rendimiento de las listas populares de canciones?
- ¿Qué atributos de canciones se correlacionan con el éxito?

Hipotesis

Hipotesis Principal

1. Las canciones más populares son recientes y tienen un alto grado de bailabillidad, bajo nivel de instrumentalidad, positivas y tienen un bajo indice de ser habladas.

Hipótesis Secundarias

2. Los cantantes en solitario y de sexo masculino y de habla inglesa son los que más destacan en canciones populares.
3. Los géneros de música que más gustan al público son el pop y el hip hop.
4. Las canciones más largas son menos populares.
5. Las canciones más ruidosas y enérgicas tienen una correlación positiva.

Descripción de los datos

El dataset es un conjunto de datos por Mark Koveha. Es un dataset publico disponible en Kaggle en el siguiente link.

Las variables contenidas en el datset son las siguientes:

- **artist**: Nombre del Artista.
- **song**: Nombre de la Pista.

- **duration_ms**: Duración de la pista en milisegundos.
- **explicit**: la letra o el contenido de una canción o un video musical contienen uno o más de los criterios que podrían considerarse ofensivos o inadecuados para los niños.
- **year**: Año de lanzamiento de la pista.
- **popularity**: cuanto mayor sea el valor, más popular será la canción.
- **danceability**: la bailabilidad describe qué tan adecuada es una pista para bailar. Un valor de 0 es menos bailable y 1 es más bailable.
- **energy**: La energía es una medida de 0 a 1 y representa una medida perceptible de intensidad y actividad.
- **key**: La clave en la que se encuentra la pista. Los números enteros se asignan a tonos utilizando la notación estándar de clase de tono. P.ej. 0 = C, 1 = C♯/D♭, 2 = D, y así sucesivamente. Si no se detectó ninguna clave, el valor es -1.
- **loudness**: El volumen general de una pista en decibelios (dB).
- mode: Mode indica la modalidad (mayor o menor) de una pista, el tipo de escala de la que se deriva su contenido melódico. Mayor está representado por 1 y menor es 0.
- **speechiness**: Speechiness detecta la presencia de palabras habladas en una pista. Cuanto más parecida a la voz sea la grabación más cerca de 1,0 será el valor del atributo.
- **acousticness**: Una medida de confianza de 0 a 1 de si la pista es acústica. 1.0 representa una alta confianza en que la pista es acústica.
- **instrumentalness**: Predice si una pista no contiene voces. Los sonidos "Ooh" y "aah" se tratan como instrumentales en este contexto. Las pistas de rap o de palabras habladas son claramente "vocales". Cuanto más cerca esté el valor de instrumentalidad de 1 mayor será la probabilidad de que la pista no contenga contenido vocal.
- **liveness**: Detecta la presencia de una audiencia en la grabación. Los valores de vivacidad más altos representan una mayor probabilidad de que la pista se interprete en vivo. Un valor superior a 0,8 proporciona una gran probabilidad de que la pista esté activa.
- **valence**: Una medida de 0.0 a 1.0 que describe la positividad musical transmitida por una pista. Las pistas con una valencia alta suenan más positivas (p. ej., felices, alegres, eufóricas), mientras que las pistas con una valencia baja suenan más negativas (p. ej., tristes, deprimidas, enfadadas).
- **tempo**: El tempo general estimado de una pista en pulsaciones por minuto (BPM). En terminología musical, el tempo es la velocidad o ritmo de una pieza dada y se deriva directamente de la duración promedio del tiempo.
- **genre**: Género de la pista.

Primeramente, realizando un análisis exploratorio de los datos, se determinó que de los 2000 registros, no existían datos nulos pero se encontraban 121 resultados duplicados que se eliminaron dejando así 1,879 registros únicos con 18 atributos en total.

A continuación podemos resumir las variables con su información correspondiente en la Tabla 1.

Tabla 1. Resumen de las variables incluidas en el dataset.

Column	Type	Non-Null	Nulls	Unique	Example
artist	object	2000	0	835	Britney Spears
song	object	2000	0	1879	Oops!...I Did It Again
duration_ms	int64	2000	0	1793	211160
explicit	bool	2000	0	1	False
year	int64	2000	0	23	2000
popularity	int64	2000	0	75	77
danceability	float64	2000	0	565	0.751
energy	float64	2000	0	580	0.834
key	int64	2000	0	11	1
loudness	float64	2000	0	1671	-5.444
mode	int64	2000	0	1	0
speechiness	float64	2000	0	837	0.0437
acousticness	float64	2000	0	1208	0.3
instrumentalness	float64	2000	0	771	1.77e-05
liveness	float64	2000	0	783	0.355
valence	float64	2000	0	760	0.894
tempo	float64	2000	0	1831	95.053
genre	object	2000	0	59	pop

Del total de las 18 variables, se tienen 14 variables numéricas y 4 variables categóricas.

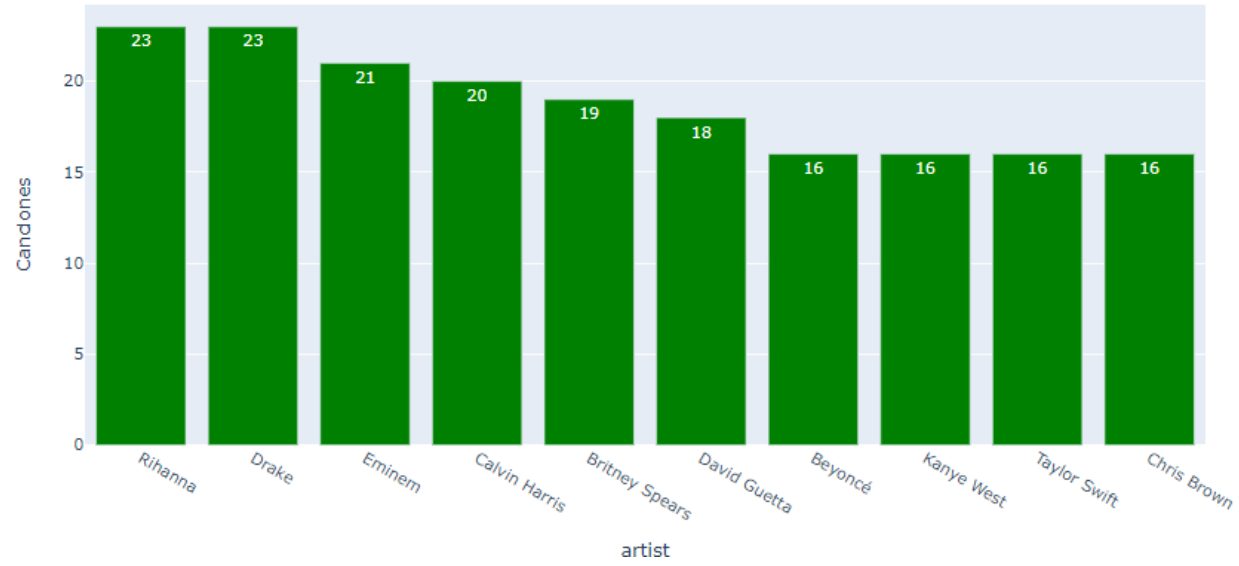
Análisis Exploratorio de Datos (EDA)

Análisis Estadístico

En el análisis exploratorio, realizamos análisis univariados y bivariados principalmente para entender lo que las variables nos dicen acerca de lo que identifica una canción al igual que cómo se relacionan todos los componentes de canciones exitosas.

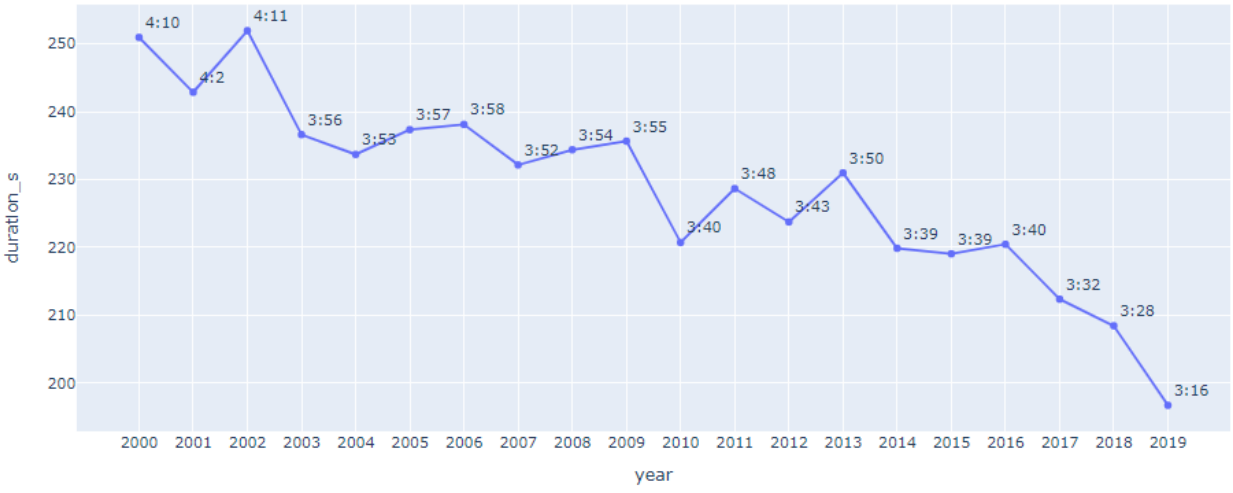
Dentro del dataframe, existen 835 artistas, de los cuales el top 10 de artistas con más canciones populares está conformado por una ligera mayoría de artistas hombres. Sin embargo es importante notar que todos estos son artistas en solitario en **Gráfica 1**.

Grafica 1. Artistas con las más canciones populares en Spotify

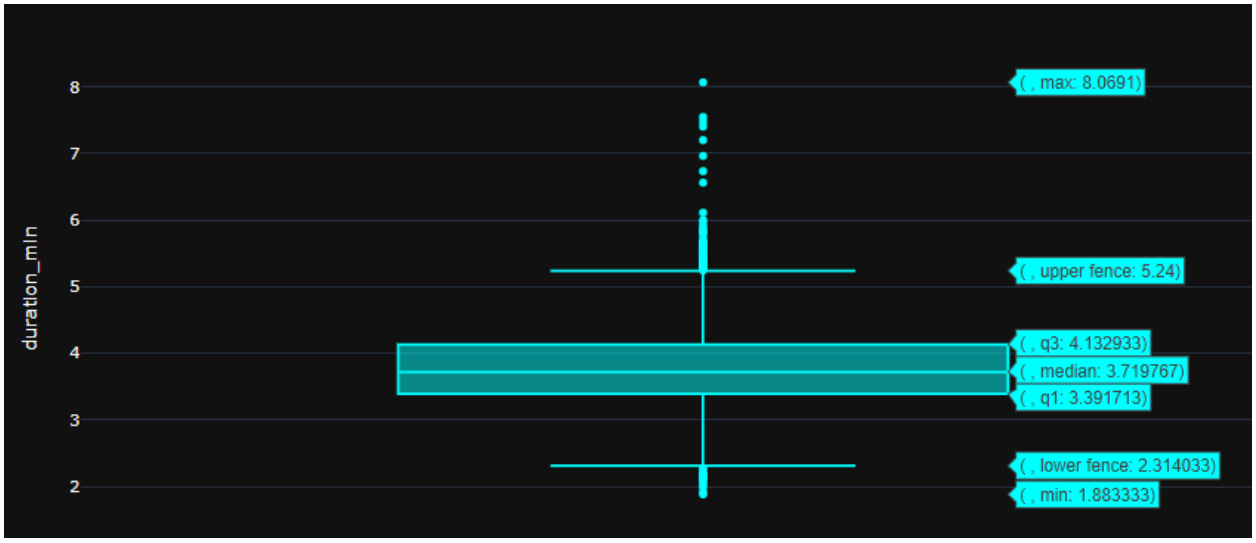


En cuestión de la duración de las canciones, el público ha ido prefiriendo canciones más cortas a través de los tiempo (**Gráfica 2**). Dentro del dataframe, el promedio de duración de las canciones es de 3.7 minutos de acuerdo a la **Gráfica 3**.

Gráfica 2. Promedio de la duración de las canciones a través de los años



Grafica 3. Duración de canciones populares

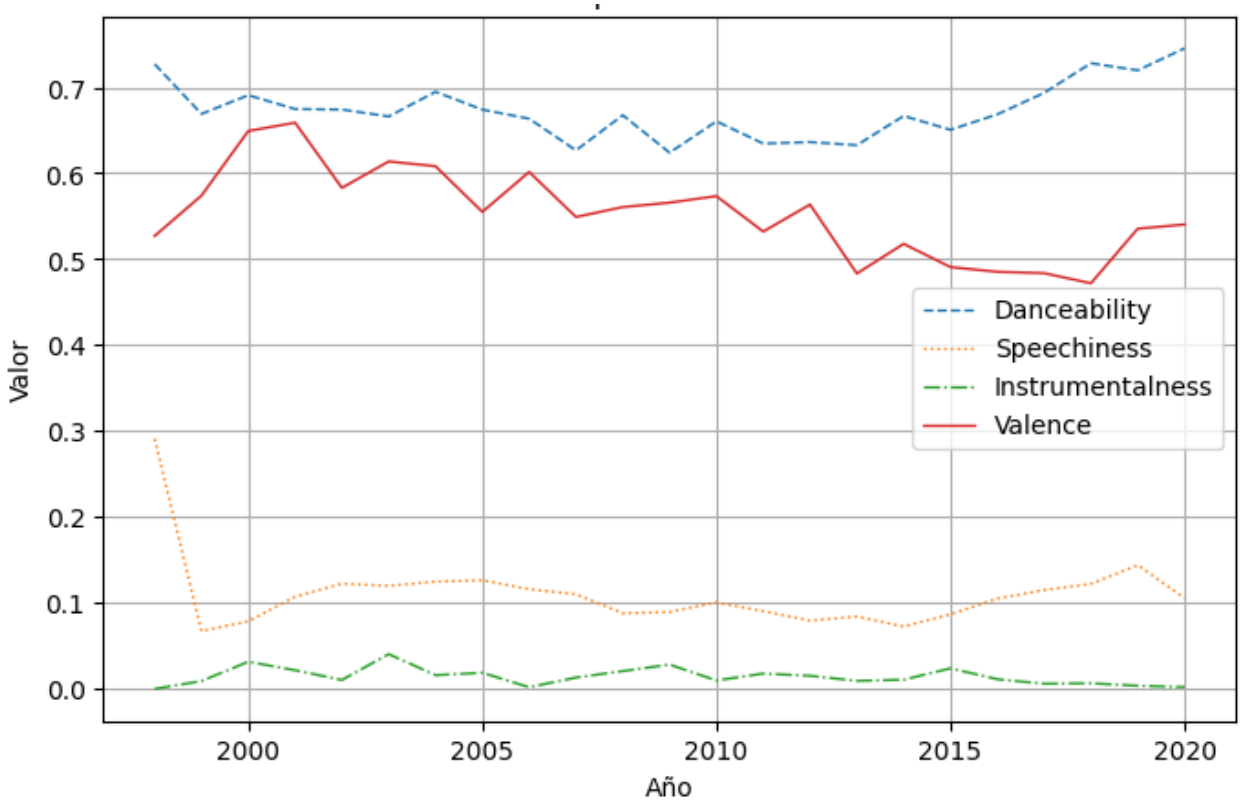


Dentro del dataframe, existen 59 géneros recopilados. Los registros del atributo género son todos los géneros detectados en cada canción por lo que el registro pop, hip hop y hip hop, pop son 3 diferentes géneros en el dataframe. A continuacion se muestra un mapa de árbol de los top 10 géneros.

Gráfica 4. Top 10 géneros musicales

Calculando los promedios de algunas de las variables de las canciones por año, se creó la **Gráfica 5**. Esta muestra algunas tendencias presentes en las variables.

Gráfica 5. Promedio de variables musicales



Los resultados son los siguientes:

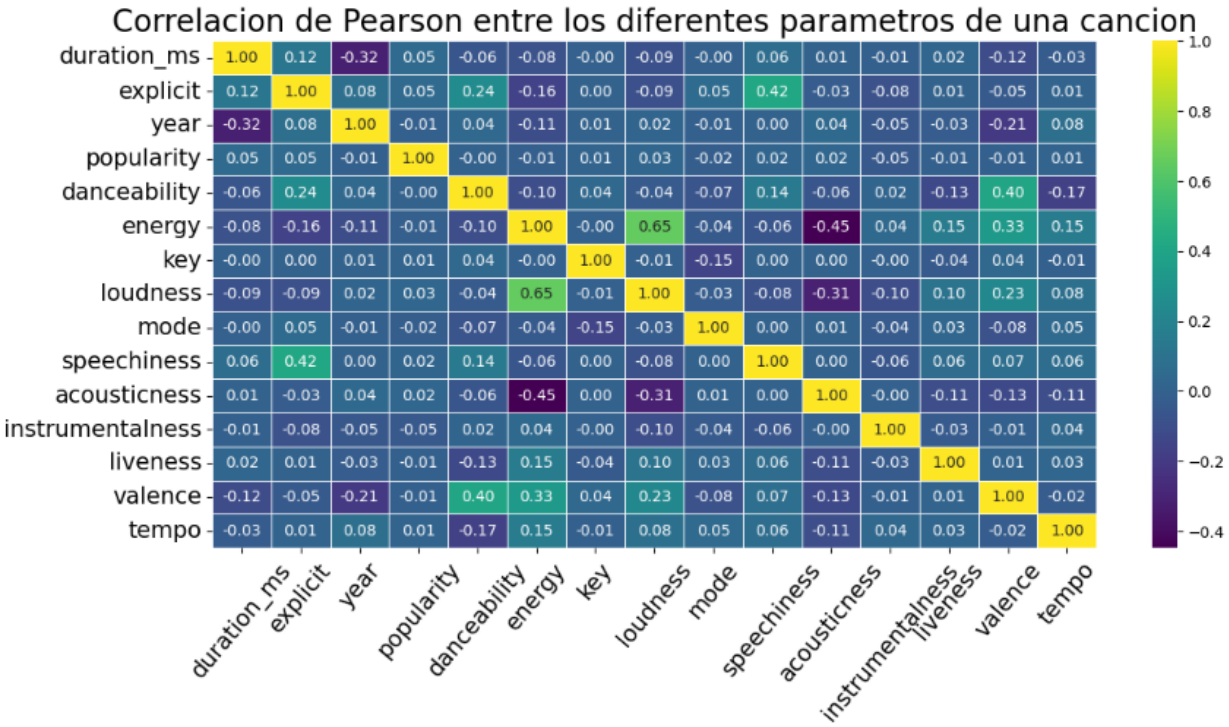
Resultados:

- La medida 'Danceability' describe qué tan adecuada es una canción para bailar. Un valor de 0.0 es menos bailable y 1.0 es más bailable. De acuerdo a la gráfica 1, la métrica sufrió una ligera disminución a través de los años. Sin embargo, a partir del 2015, está empieza a tener una tendencia creciente.
- La medida 'Valence' describe la positividad musical transmitida por una canción de 0.0 a 1.0. Las pistas con una valencia alta suenan más positivas (p. ej., felices, alegres, eufóricas), mientras que las pistas con una valencia baja suenan más negativas (p. ej., tristes, deprimidas, de enojo). De acuerdo a la gráfica, se observa una tendencia decreciente a partir de los inicios de los 2000s para luego iniciar un repunte alrededor del 2018 en camino al 2020. Es interesante notar que esta es la medida con una tendencia más fuerte y da a entender que las canciones empezaron a connotar temás más negativos a través de los años.
- La medida 'Speechiness' detecta la presencia de palabras habladas en una pista. Cuanto más parecida a una conversación sea la grabación (por ejemplo, programa de entrevistas, audiolibro, poesía), más cerca de 1.0 será el valor del atributo. La gráfica nos muestra un desplome significativo a finales de los 90s, lo cual puede ser explicado por la transición de popularidad del rap y hip hop a canciones con sintetizadores, batería electro-punk, reverb, géneros dubstep y EDM (música electrónica). Lo cual hace que establezca estabilidad a través de los años.
- Finalmente, la medida 'Instrumentalness' representa la probabilidad de que la canción no contenga contenido vocal según el valor se acerque más a 1.0. Esta es la métrica

más estable de las presentadas en el gráfico. La tendencia de los valores son valores menores por lo que indica que las canciones populares siguen siendo más los que la gente puede cantar las letras.

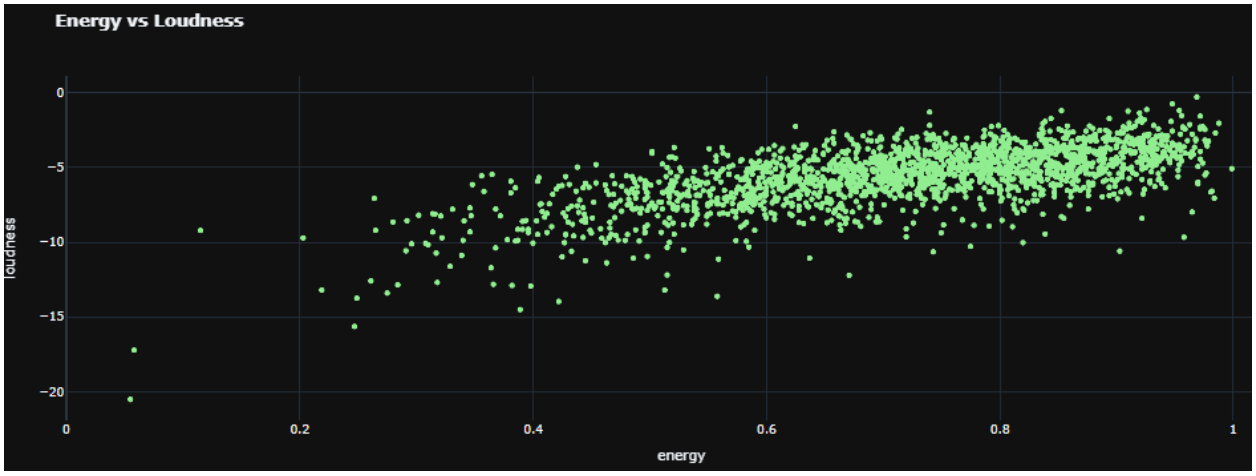
Debido a la gran variedad de variables y su naturaleza, no existe mucha correlación entre las variables donde se pueda determinar que una variable es dependiente de otra según **Gráfico 6** donde se grafica la correlación de todas las variables.

Gráfico 6. Correlación de Pearson entre los diferentes parámetros de una canción



La mayor correlación es aquella entre energy y loudness de una canción la cual graficada se ve de la siguiente manera:

Gráfico 7. Energy vs loudness



Aplicación de Algoritmos

Preparación

Para preparar el dataset para la aplicación de algoritmos, debemos transformar los datos en torno a la columna de género. Primeramente, filtramos los datos por los géneros más relevantes es decir creamos un dataframe donde solo se encuentren las canciones más populares cuyo género ha tenido una frecuencia en el dataframe de al menos 10 veces.

Posteriormente, el atributo género es de tipo variable categorial por lo que se les asignó un valor numérico, mediante el algoritmo de ***labelencoder*** para poder hacerlo viable para un modelo de machine learning que nos permita predecir.

Modelos y métricas de desempeño

Los modelos utilizados para predecir la variable ***genre*** fueron Random Forest y KNN. Se testean los modelos con el nuevo dataset creado, filtrando las variables que se mencionaron en la sección anterior, tomando una división de Train/Test con una relación de 80/20. Para posteriormente implementar una optimización del mejor modelo según las métricas de accuracy, precision y recall.

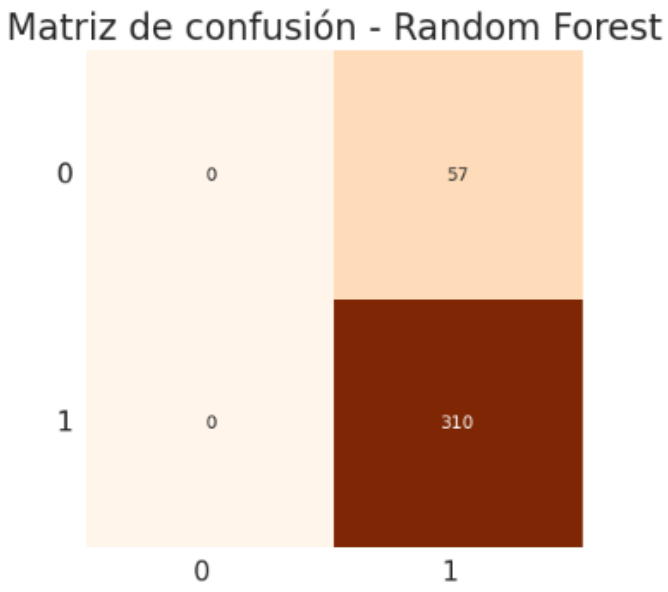
A continuación se presentan los desarrollos de los modelos:

Random Forest

Se aplicó un número de 200 estimadores para este modelo y random state de 11. Con esto se calculó el accuracy del set de entrenamiento y de evaluación, cuyos resultados fueron 0.9993 y 0.8446 respectivamente.

Para la validación del modelo se decidió aplicar el cross validation, el puntaje fue de 0.8181. También su matriz de confusión se presenta a continuación:

Gráfico 8. Matriz de confusion de Random Forest



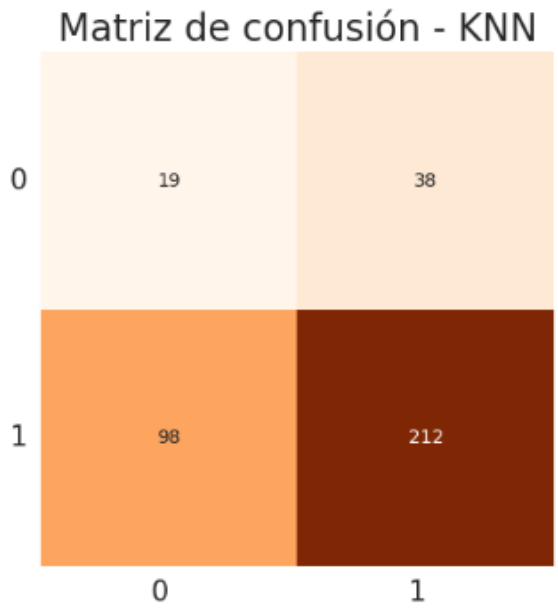
K-Nearest-Neighbor (knn):

Se buscó hacer la predicción de género musical por medio del comportamiento de los vecinos más cercanos también, en el espacio multidimensional compuesto por las demás variables, usando 2 vecinos mas cercanos. Esto debido a que existe una enorme variación y reiteración de datos que la mayor proximidad debería de ser determinante para predecir lo que se busca.

Con esto se calculó el accuracy del set de entrenamiento y de evaluación, cuyos resultados fueron 0.86 y 0.63 respectivamente.

Para la validación del modelo se decidió aplicar el cross validation, el puntaje fue de 0.61. También su matriz de confusión se presenta a continuación:

Gráfico 9. Matriz de confusión del modelo KNN



Finalmente, se calculan las métricas accuracy, precisión, recall y AUC presentadas en Tabla 2.

Tabla 2. Métricas de desempeño de Random Forest y KNN

Modelo	Accuracy	Precision	Recall	AUC
`Random Forest`	0.84	0.84	1.00	0.50
`KNN`	0.63	0.85	0.68	0.51

De acuerdo a las métricas el Random Forest resulta ser el mejor modelo a elegir para este caso.

Optimizacion

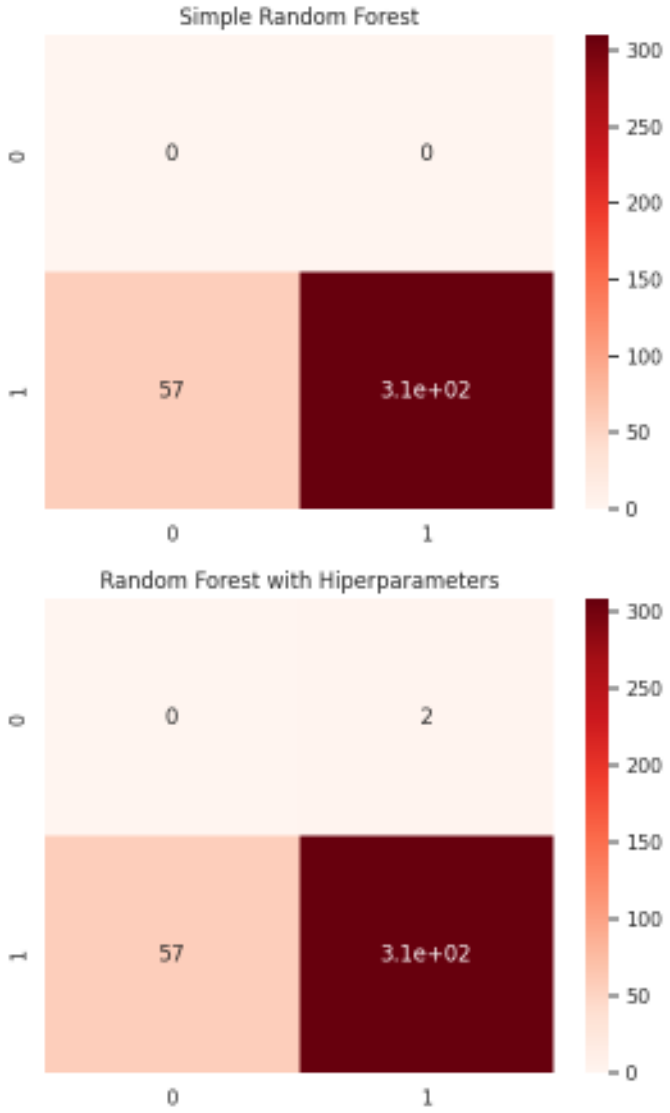
Hyperparameter Optimization: Partiendo del modelo que mejor se adapta a los datos (Random Forest) intentaremos ajustar sus parámetros para conocer si es posible mejorar su precisión.

Se utilizó el método de Grid Search (búsqueda por grilla) para optimizar los hiperparametros del Random Forest:

```
# Numero de Árboles del Random Forest
n_estimators = [int(x) for x in np.linspace(start = 5, stop = 20, num =5)]
# Numero de caracterísiticas a considerar en cada split
```

```
max_features = ['auto', 'sqrt']
# Número máximo de niveles por Arbol.
max_depth = [2,5]
# Número mínimo de muestras necesarias para dividir un nodo.
min_samples_split = [2,5]
# Número mínimo de muestras necesarias en cada nodo hoja.
min_samples_leaf = [2,5]
# Método de selección de muestras para entrenar cada árbol.
bootstrap = [True, False]
```

Los resultados nos arrojaron las siguientes matrices de confusion:



Con esto se puede notar que la diferencia no es significativa por lo que la optimización ya se ha alcanzado.

Recomendaciones

Parte de la gran complejidad del proyecto recae en la gran variabilidad de los elementos que puede contener una canción. Esto fácilmente se refleja en el manejo y transformación del atributo de género musical ya que la manera en que se recopilan estos datos y se estudia el arte de la música es que una canción puede clasificarse con una combinación de géneros. La limitante presentada en el modelado de este caso es la predicción de si una canción pertenece solo al género pop o al género hip hop, pop. Quizás la solución sería alimentar de más variables precisas a un modelo para una entrenar una correcta predicción sin embargo se presenta como un reto para la siguiente vez.

Asimismo otra limitante es el origen de la recopilación de datos ya que Spotify es una plataforma global. El top de canciones populares sin duda puede variar entre diferentes países y regiones por lo que esta base de datos es una gran generalización. Sería importante considerar esto tanto en los registros únicos de canciones como en el cálculo de la variable popularidad, la cual no fue del todo claro cómo se determinaron los resultados de este atributo en la base de datos.

Finalmente, se presenta una gran oportunidad de profundizar la información de las tendencias y promedios de canciones populares aquí estudiada si se complementa con algún estudio de los contextos históricos de las fechas de lanzamientos y de fechas donde las canciones alcanzaron una mayor popularidad. Por ejemplo, el conocer el contexto histórico de cuándo una canción en el 2014 fue popular por la alza de artistas de música electrónica puede explicar la baja en la variable speechiness que mide el grado de palabras pronunciadas en una canción.

Conclusión

La receta de un hit musical en Spotify involucra una infinidad de variables que se vuelve difícil determinar las variables enteramente dependientes de otras. Sin embargo, el estudio de esta materia puede resultar altamente beneficioso para el estudio social de las preferencias del público a través de los años.

En conclusión, los insights que se pudieron recopilar sobre las canciones más populares son que los géneros que más predominan y posiblemente predominarán son el pop y el hip hop. Es por esto que los artistas tienden a ser cantantes y raperos, interesantemente en solitario. Esto

también nos explica que para crear una cancion popular con el público debe contener vocales para que sea mas fácil cantar.

Para alcanzar una mayor audiencia, la canción debe ser en el idioma de inglés y en general no sobrepasar la marca de 4 minutos en duración. Adicional, gustan más las canciones que se pueden bailar facilmente y los temas que no sean del todo positvos.