UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Informàtica

etsinf

LA MÚSICA EN PERSPECTIVA:

Un análisis de las canciones de hoy para entender el mañana

Proyecto II. Integración y preparación de datos

Grado en Ciencia de Datos 2020-2021

Carlos Gallego Andreu
Alejandro Losa Brito
Héctor Martínez Cabanes
Daniel Oliver Belando
Daniel Romero Alvarado

Presentación

Este trabajo ha sido desenvuelto en la asignatura de Proyecto II del grado de Ciencia de Datos de la Universidad Politécnica de Valencia. A partir de un tema general, desarrollaremos todas las fases de un proyecto, plantearemos objetivos y aplicaremos conocimientos adquiridos durante los 2 primeros años del curso, centrando el análisis en las técnicas aprendidas en el segundo curso del grado. El tema de nuestro proyecto es la **Música.**

Introducción

Es indudable que la música forma una parte muy importante de nuestra vida, siempre ha sido una forma de expresar nuestros sentimientos y de unir socialmente a la gente, en todo tipo de festividades y celebraciones. La gran variedad de géneros hace que cada persona tenga unos gustos musicales únicos, que incluso llegan a formar parte de su personalidad y su identidad.

Gracias a la revolución tecnológica de los últimos 20 años, podemos acceder a mucha información sobre las canciones online. Aunque no sepamos la cantidad exacta de canciones que hay en total, es seguro que no faltan, a través de plataformas como Spotify, YouTube o iTunes, se suben casi 60.000 canciones al día. De hecho, Spotify es una de las plataformas con más canciones ofertadas del mundo (unos 30 millones en 2018).

En este proyecto, mostraremos como nuestro equipo ha aprovechado toda esta información, desde la integración de distintas fuentes de datos, hasta la extracción de conclusiones de estos, para llevar a cabo un estudio con distintos objetivos de interés investigativo y comercial.

Agradecimientos

A nuestras familias, por el apoyo brindado durante todo el proyecto: en momentos de frustración y desasosiego, siempre han estado a nuestro lado.

A María José Ramírez Quintana y a Sara Blanc Clavero por habernos orientado a lo largo del proyecto, perfilando las decisiones que hemos tomado y animándonos a seguir adelante.

A los usuarios de Kaggle Yamac Eren Ay y Rodolfo Figueroa, por haber proporcionado gran parte de los datos para el estudio.

Notas

Todas las imágenes presentadas en este documento han sido elaboradas por el equipo mediante el uso de software de visualización de datos, principalmente RStudio.

Índice de contenidos

1. Alcance del proyecto

El proyecto que hemos planteado se divide en 3 partes únicas e indispensables, ya que cada una de ellas aporta elementos importantes a nuestro estudio: por un lado, la centralización de varias fuentes de datos de canciones y sus características técnicas y musicales, tanto ya creadas como de elaboración propia, en un solo archivo; por otro lado, un estudio de las características y de las relaciones entre ellas; y por último, y a través de distintos modelos de clasificación supervisados y no supervisados, las conclusiones extraídas de la base de datos para cada objetivo que planteamos.

1.1. Objetivos del proyecto

El objetivo principal de este proyecto es elaborar un sistema de recomendación de canciones a partir de la información (25 variables) de 1.300.000 canciones de Spotify basado principalmente en recomendar canciones a un usuario que tengan características similares (principalmente acústicas, como tempo, sonido, compás, pero también otras como la popularidad) a las que ya ha escuchado. ¿Vale la pena recomendar canciones similares, pero poco populares, o es mejor recomendar canciones populares, aunque guarden menos similitudes?

Otro de los objetivos, quizá de carácter más investigativo, es ver la evolución de las características de las canciones a lo largo de los años, ¿por qué algunas de ellas siguen siendo muy populares? ¿Guardan las canciones populares alguna característica similar, que se mantiene a lo largo de los años? O incluso, ¿Es posible determinar si una canción será popular o no, en base a sus características musicales?

Tras el planteamiento del primer objetivo, surge uno general ¿podemos clasificar canciones similares, en base a sus características musicales?, es decir, ¿las variables que tenemos son suficientes para distinguir perfectamente distintas clases de canciones? Puede parecer redundante, pero al tener tanta variedad de canciones, no sabemos el grado de dificultad al que nos enfrentamos.

1.2. Utilidad del estudio

El primero de los objetivos puede tener interés comercial: con un buen sistema de recomendación podemos dar a conocer canciones que el usuario no conocía, pero que son similares. Esto, aunque grandes empresas como Spotify o ITunes ya lo tengan implementado, puede ser interesante para otras más pequeñas, como Audius.

El segundo objetivo puede tener más interés para los propios artistas: si, históricamente, las canciones populares llegaran a guardar alguna relación en cuanto a alguna característica musical, podrían implementarla en sus canciones y aumentar las probabilidades de llegar a más gente. Un ejemplo de esto podría ser el ritmo del reggaetón, presente en prácticamente todas las canciones populares de este género latino.

2. Técnicas utilizadas

Durante las distintas etapas del proyecto nos hemos apoyado en herramientas de procesamiento de datos muy diferentes. Durante la limpieza e integración de las fuentes de información, utilizamos el lenguaje **Python** y las librerías *Spotipy*, *requests* y *pandas* para acceder a una API REST[1], y completar información que no teníamos disponible sobre varias variables de interés. También utilizamos **RStudio** para combinar 2 archivos y así tener más datos disponibles para el estudio.

[1] Una API REST es una interaz específica de un servicio (como Spotify) para obtener datos de su plataforma que pretende facilitar el acceso a dichos datos

Por último, los análisis y las conclusiones se obtuvieron a partir de RStudio mediante técnicas de *clustering*, *PCA* o *análisis discriminante*. Estos análisis se explicarán con más detalle en el apartado 4.

El equipo también ha hecho uso de otras herramientas de carácter transversal a lo largo del proyecto, que han ayudado a la organización y comunicación entre los miembros: entre ellos, **Google Drive**, para tener disponible la información de cada uno en todo momento, así como para la organización de las distintas fases del estudio; **Trello** para organizar objetivos a corto y largo plazo, y asignar las tareas a cada miembro del equipo; y **Deepnote** para trabajar en Python al mismo tiempo. También se realizaban 2 reuniones semanales en **Microsoft Teams** para llevar el proyecto al día, una de ellas con una profesora de la asignatura, en la que nos solucionaba dudas y nos daba ideas para conseguir nuestros objetivos. Por último, utilizamos **iMovie** para realizar el vídeo explicativo del proyecto.

3. Fases del proyecto

En este apartado explicaremos las tareas que realizamos antes de responder a los objetivos planteados; hablaremos de las fuentes de datos, el procesado, la limpieza y el análisis exploratorio.

Para hacer la memoria más amena, la mayor parte de la información sobre estos temas estará disponible en el anexo, aquí solo mostraremos lo más importante e indispensable para el proyecto.

3.1. Fuentes de datos

Para seleccionar las fuentes de datos hemos seguido varios criterios, como, por ejemplo, los atributos de cada base de datos y cómo pueden contribuir a nuestros objetivos. De esta manera, hemos seleccionado aquellos conjuntos de datos que describen las canciones con características sobre los artistas, la popularidad y varias relacionadas con el ritmo o el sonido.

Finalmente se seleccionaron dos fuentes de datos, con información técnica y musical sobre las canciones. Ambas bases de datos fueron extraídas de **Kaggle**[2], y contenían canciones obtenidas de Spotify mediante una API. Gracias al identificador de la canción pudimos comprobar si estaban repetidas, e hizo más fácil la posterior integración en un solo archivo.

Una de las fuentes es Spotify_dataset_1921-2020, de 172,230 canciones entre 1921 y 2020, y la otra es Spotify 1.2M+ songs, de 1,204,025 canciones. Consideramos que estas bases de datos eran las más completas y las que más información nos daban sobre canciones. Al tener muchas variables numéricas podemos hacer estudios como PCA o clustering, para encontrar grupos o relaciones.

Cabe destacar que dos variables que consideramos muy importantes, 'popularity' y 'time_signature' no estaban en este segundo conjunto, y tuvimos que completarlas mediante la API de Spotify.

Se puede encontrar información más detallada de las bases de datos empleadas en el Anexo I.

[2] Kaggle es una comunidad en línea de científicos de datos y profesionales del aprendizaje automático que permite encontrar y subir conjuntos de datos de manera pública

3.2. Integración y transformación de los datos

En este apartado se explicará cómo combinamos las 2 fuentes de datos, cómo rellenamos los valores faltantes y cómo retocamos y añadimos otras variables para completar el fichero con el que trabajamos.

La unión de las bases de datos se hizo mediante RStudio, y se encontraron algunos atributos con valores faltantes ya que ambos datasets no tenían las mismas variables. Para solucionarlo empleamos la API de Spotify, que nos permitía acceder a cualquier característica de la canción.

3.2.1. Transformación de variables: artistas

Algunas canciones estaban creadas por varios artistas en conjunto, así que decidimos quedarnos con el artista principal para que las conclusiones que obtuviéramos en los análisis fueran más claras y fáciles de interpretar. Tampoco podríamos haber descubierto la gran cantidad de artistas clásicos (se comenta en el siguiente apartado) sin realizar este cambio, por lo que ha sido importante para el estudio.

Además, los artistas se encontraban como texto entre corchetes (es decir, no era una lista, sino una cadena de caracteres), disminuyendo su legibilidad y utilidad, por lo que realizamos los cambios correspondientes mediante expresiones regulares (una herramienta para identificar patrones en cadenas de texto) para modificar este formato en la variable.

Se puede encontrar una descripción detallada del proceso y el código empleado en el Anexo III.

3.2.2. Completando la información con Spotipy

Spotipy es una librería de Python que nos permite obtener información sobre una canción que esté en la base de datos de Spotify. Podemos acceder a ella mediante su ID o buscando el álbum o la *playlist* donde se encuentra.

Gracias a que contamos con los identificadores de las canciones en nuestros datos, podemos consultar sus valores de popularidad y compás con la API, información que no estaba en una de las fuentes de datos y que es importante para nuestro estudio.

Mediante la API también añadimos una nueva variable a la base de datos: 'classical', una variable binaria que nos permitía ver si una canción era de género clásico o no. Al hacer un análisis inicial, nos dimos cuenta de que existía una mayoría de canciones de artistas clásicos (Beethoven, Mozart, Bach, etc.) que distorsionaban los resultados, y acabamos dividiendo el dataset en canciones clásicas y no clásicas. Nuestros análisis se centrarán en las canciones no clásicas.

Se puede encontrar una descripción detallada del proceso y el código empleado en los Anexos IV y VV.

3.3. Análisis exploratorio

En este apartado mostraremos los primeros análisis que hicimos sobre la base de datos, que tienen como objetivo familiarizarnos con las variables y conocerlas mejor.

3.3.1. Información sobre las variables

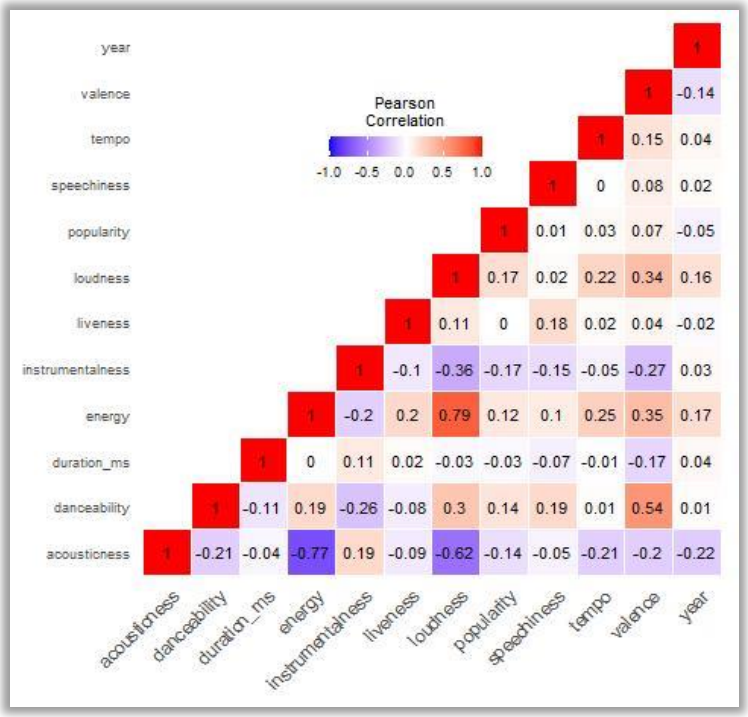
Por una parte, tenemos variables categóricas muy variadas como el nombre del **artista**, de la **canción**, o del **álbum**, con sus correspondientes identificadores **ID** de Spotify. Por otra parte, contamos con muchas variables numéricas y binarias, que se pueden dividir en distintos temas:

- El estado de ánimo: cómo de **bailable** (danceability), **positiva** (valence) o **enérgica** (energy) es una canción.
- Propiedades técnicas: cómo de **ruidosa** es (loudness), si tiene muchos **instrumentos** (instrumentalness), o en cambio la mayoría de la música es **cantada.**
- Información sobre el contexto: por ejemplo, detecta la presencia de una **audiencia** (liveness) en la canción (por si hubiera sido grabada en un concierto).
- Otra información: como la **duración** (duration_ms), el **año** de salida (year), o si la canción tiene o no letra **explícita** (letra considerada ofensiva, inadecuada para niños, lenguaje violento o lenguaje discriminatorio).

Se puede encontrar una descripción detallada de las variables empleadas en el estudio en el Anexo II.

3.3.2. Análisis de canciones no clásicas

Dado que contamos con las canciones clásicas y no clásicas por separado se decidió analizar solo las canciones de artistas no clásicos y serán las que tendremos en cuenta para los objetivos.



Como podemos observar acousticness está relacionado negativamente con energy y loudness y estas dos últimas están correlacionadas positivamente, por lo que a mayor volumen tenga la canción(loudness) mayor será la energía de esta. Además, simplificando lo dicho al principio, cuanto más acústica sea la canción menor volumen y menor energía tendrá.

Figura 1. Matriz de correlaciones de las variables numéricas de las canciones no clásicas

Aunque el coeficiente de correlación es menor, podemos ver que a mayor danzabilidad tiene la canción mayor positividad transmite (valence). También sorprende que la popularidad esté tan poco correlacionada con las demás, lo que dificulta nuestro segundo objetivo.

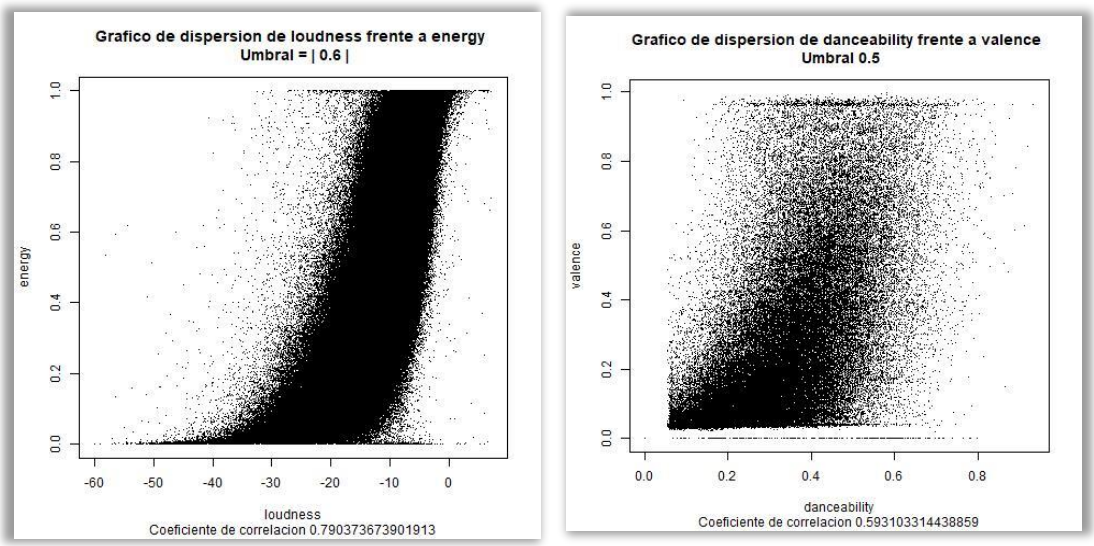


Figura 2. A la izquierda, gráfico de dispersión de volumen frente a energía. A la derecha, gráfico de dispersión de danzabilidad frente a valencia.

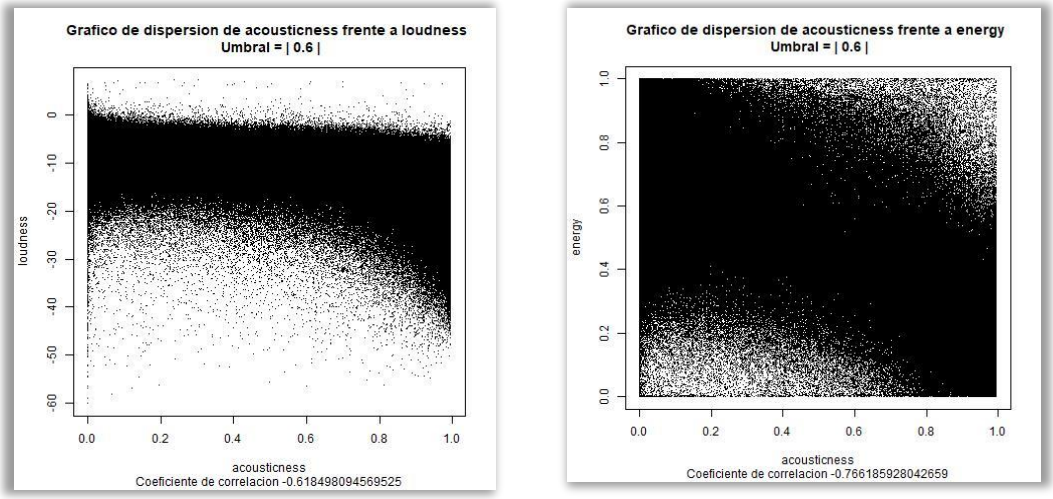
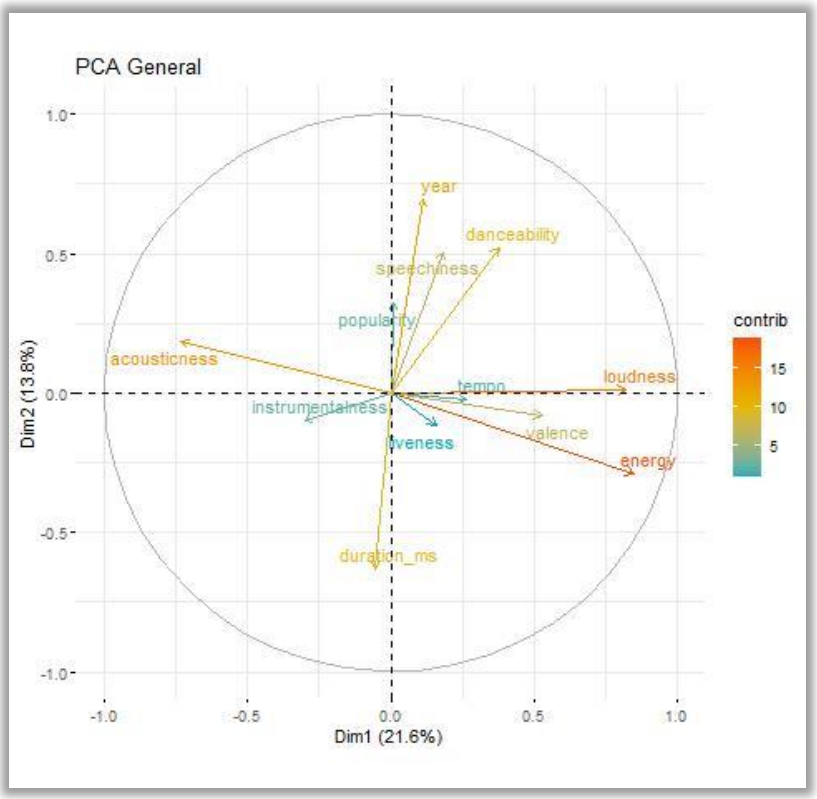


Figura 3. A la izquierda, gráfico de dispersión de acústica frente a volumen. A la derecha, gráfico de dispersión de acústica frente a energía

Se puede encontrar el módulo empleado para generar estos gráficos en el Anexo VI.

Decidimos completar el análisis exploratorio con un análisis de componentes principales, que nos permite ver las relaciones entre las variables, y en relación con las propias canciones.



Al tener bastantes variables en el estudio y presentar poca correlación entre ellas, las primeras dimensiones solo explican un 35,4%. Aun así, podemos ver que cuanto más energética, más ruido tiene, o que cuanto más bailable y cantada es una canción, más reciente será. También es curioso ver que, a mayor duración, más antigua es la canción.

Figura 4. Gráfico de loadings del PCA realizado a nuestros datos

4. Resultados obtenidos

En este apartado mostraremos los estudios y las conclusiones correspondientes de cada objetivo: ¿Qué características tienen las canciones a lo largo de los años? ¿Podemos predecir la popularidad que tendrá una canción, a través de su información musical? Y, por último, ¿cobrará relevancia la construcción de un sistema de recomendación?

4.1. Las canciones a lo largo de los años

Muchas canciones que se lanzaron hace más de 60 años siguen siendo populares hoy en día. ¿tienen estas canciones las mismas características musicales que las canciones actuales?

Para este estudio hemos realizado un Análisis de Componentes Principales (PCA), filtrando los datos de las canciones por décadas. El PCA nos permite ver relaciones entre variables, y saber qué características tienen más peso dentro de un conjunto de datos. Las librerías de R utilizadas han sido *FactoMiner*, *factoextra* y *dplyr.*

4.1.1. Estudio

En nuestra base de datos disponemos de muchas más canciones actuales que antiguas. Dada nuestra falta de capacidad computacional para realizar PCAs de más de 30,000 canciones, en las últimas décadas tuvimos que filtrar los datos, eliminando canciones con popularidad < 10 o < 20, viendo la distribución de la variable popularidad (ver Anexo VII).

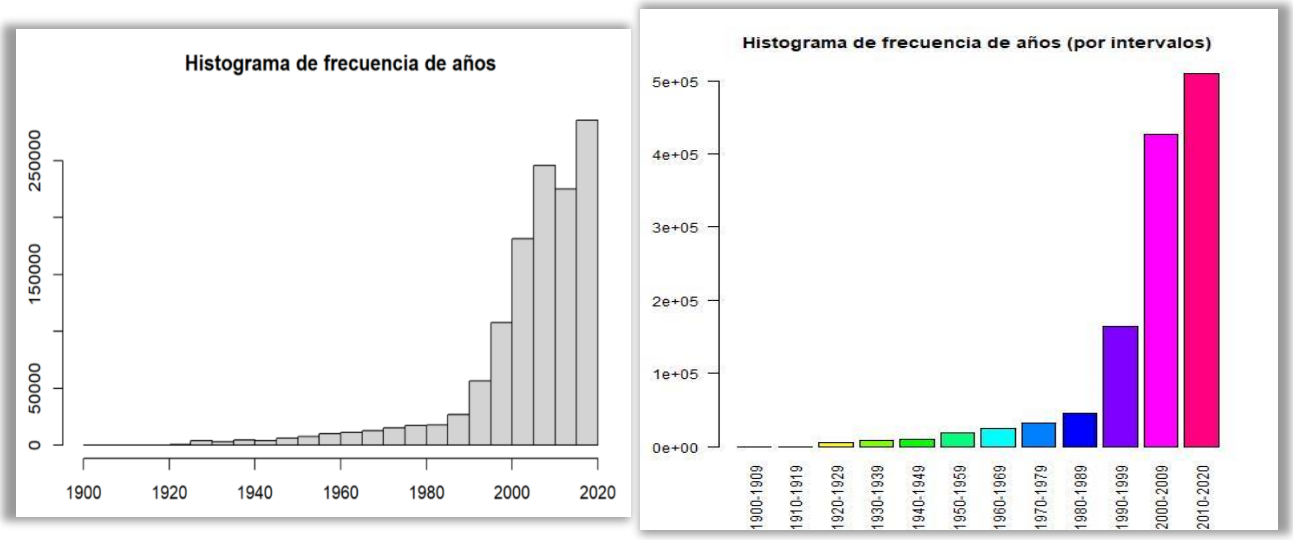


Figura 5. A la izquierda, histograma de frecuencia de la variable year. A la derecha, diagrama de barras de la variable year_interval, una variable qua agrupa los años por décadas.

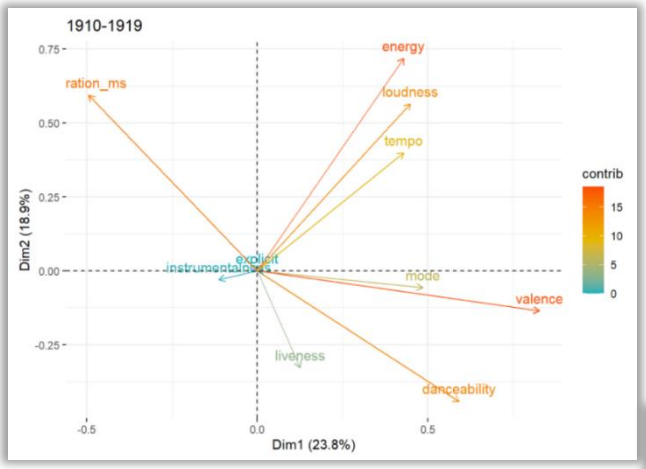


Figura 6. Gráfico de loadings de la década 1910-1920

Empezando en 1910s, la **positividad** es la característica que más destaca. El **tempo** y la **clave** son otras variables a tener en cuenta. En esta década se favorecen canciones **alegres**, **rápidas**, **cortas** y que transmiten una sensación muy positiva.

A medida que pasan los años, la **energía**, el **volumen**, la **positividad** y la **danzabilidad** emergen como características importantes. Otras variables como la **duración**, **liveness**, o la falta de **instrumentalidad** también se ven favorecidas en esta década en concreto.

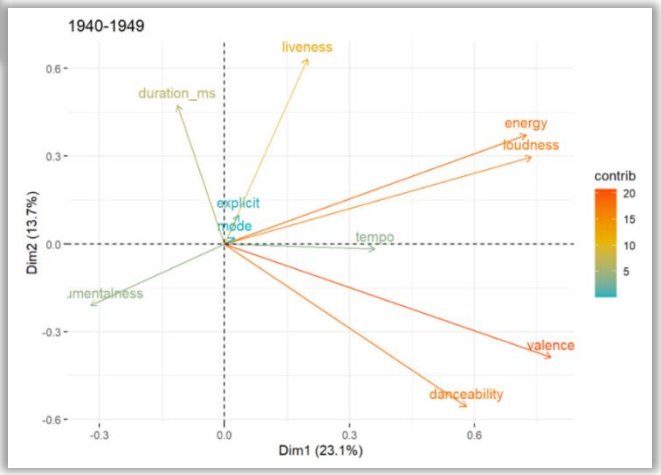
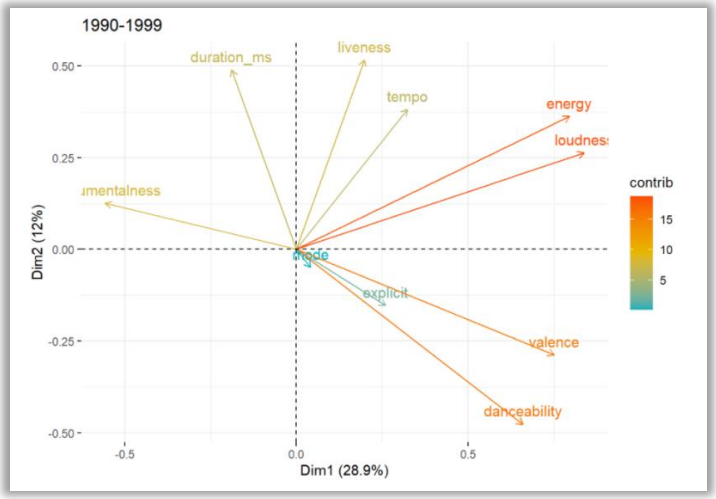


Figura 7. Gráfico de loadings de la década 1940-1950



En los 1990s las canciones **rítmicas** y por primera vez, con letra **explícita**, parecen tener importancia. Pero el esquema principal de los últimos 50 años es el mismo, las canciones **enérgicas**, **bailables** y **positivas** son las características claves en las canciones.

Figura 8. Gráfico de loadings de la década 1990 – 2000

En las últimas 3 décadas tampoco hay mucha diferencia entre las variables: los cambios más radicales vienen en la característica '**explicit'**, llegando incluso a tener la misma importancia que la **positividad**, algo que no había ocurrido nunca.

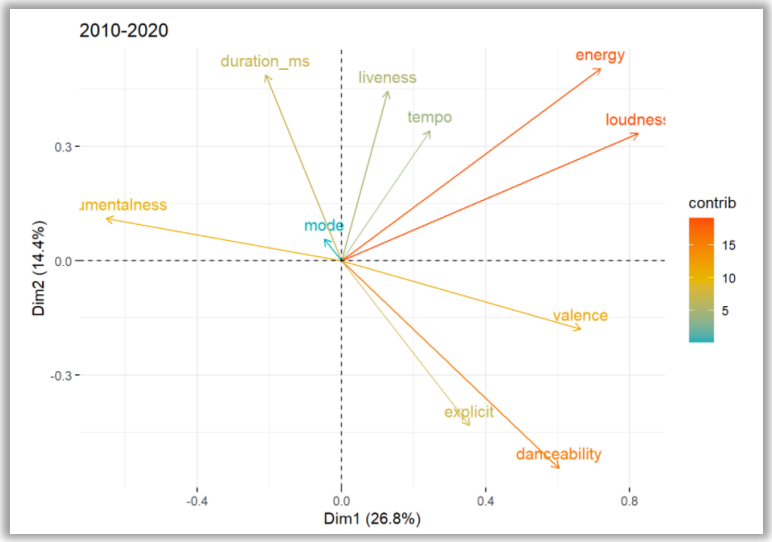


Figura 9. Gráfico de loadings de la década 2010 – 2020

Pese a existir pequeñas diferencias entre cada década, hay una preferencia universal a lo largo del tiempo: canciones rápidas, de alto volumen, que puedan ser bailables y que hagan felices a quienes las oigan. Las mayores diferencias parecen estar en variables más sutiles de las canciones, como la duración, la aparición de contenido explícito o el ritmo de la canción. En conclusión: la diferencia está en los detalles.

Por cuestiones de espacio hemos decidido seleccionar cuatro gráficos de loadings (los más representativos del período), pero hemos realizado un gráfico de loadings para cada década. Todos estos gráficos, así como su correspondiente interpretación, pueden ser encontrados en el Anexo VII.

4.2. Predicción de la popularidad a través de atributos musicales

¿A qué artista no le gustaría saber si su canción será un éxito o no? La popularidad de una canción valora la cantidad de reproducciones de una canción (en Spotify) durante un cierto período de tiempo.

Este parte tiene fines predictivos, no obstante, la cuantificación de la contribución de cada variable para obtener la popularidad también es uno de los principales objetivos en este estudio.

Partiendo de las características musicales, intentaremos determinar el valor de la popularidad, por una parte, realizando una regresión lineal múltiple (MLR), intentando predecir el valor exacto de la popularidad, y, por otra parte, un análisis discriminante, buscando predecir intervalos de esta variable respuesta.

4.2.1. Estudio

Antes de empezar, casi 500,000 canciones tienen popularidad 0, por lo que cualquier modelo se podría ver influido a obtener valores bajos de popularidad. En consecuencia, no tendremos en cuenta estas canciones para normalizar la distribución de la variable y obtener mejores resultados. Tampoco tendremos en cuenta el año de salida de la canción, porque no queremos que sea un factor que influya a predecir la popularidad. Esto es debido a que la intención es predecir utilizando variables que representen características musicales de la canción y las conclusiones que obtengamos serán independientes de la fecha de salida de esta.

Regresión Lineal:

El proceso seguido para la obtención de un modelo óptimo ha estado basado en la experiencia que nos ha proporcionado el método prueba y error: es decir, hemos valorado varios modelos que según lo aprendido en análisis previos considerábamos que podrían resultar eficaces.

Antes de entrar en la creación de modelos, decidimos generar varios gráficos como el de correlaciones, importante para evitar modelos con problemas de multicolinealidad, y gráficos de las distribuciones de las variables por tal de saber si pertenecen a una distribución normal o no.

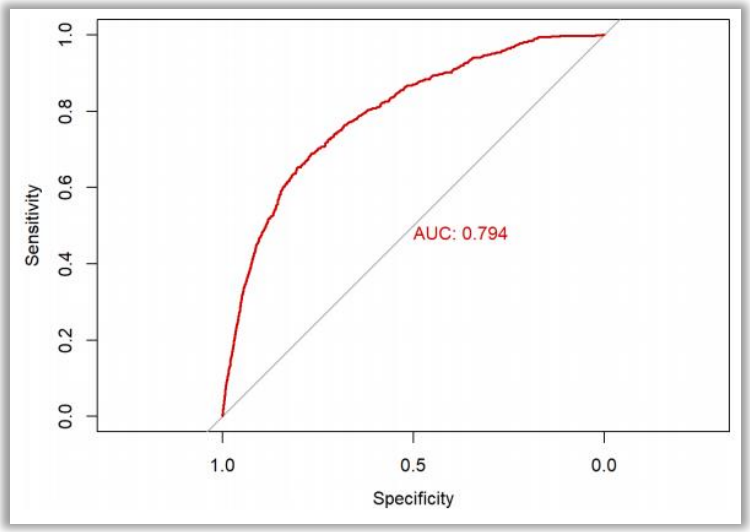
A pesar de saber de los problemas que podría llevar crear un modelo juntando todas las variables, decidimos hacerlo y saber qué resultados obteníamos. En ellos, destacamos que se explica tan sólo un 5'6% de la variabilidad en los datos, aspecto que nos indica que un modelo de regresión no es buena idea para la predicción con los datos que estamos utilizando y, además, nos introduce a la idea de que los datos que tenemos no son suficientes para predecir la popularidad.

A continuación, pasamos a crear modelos tomando menos variables, pero que cobran sentido al análisis y al modelo. Decidimos fijarnos en la matriz de correlaciones y al Análisis de Componentes Principales para seleccionar variables a añadir que no estén relacionadas entre ellas. Además, en la selección de variables valoramos que estas sean las que más contribuyen a explicar las dimensiones y que mejor ayudan a explicar el conjunto de datos. No obstante, como era de esperar los resultados muestran un valor más bajo que el mostrado con el modelo que incluía todas las variables, deduciendo de esto que las características musicales resultan ser insuficientes para predecir la popularidad que tendrá una canción. Cabe añadir que sería interesante crear un modelo añadiendo aspectos como, por ejemplo, el año de la canción o el artista, ya que por sentido común y experiencia sabemos que las canciones más recientes suelen ser habitualmente las más populares en la actualidad y los datos recogidos en esta variable son actualizados en referencia a las reproducciones de las canciones en un periodo cercano al presente. Además, pensamos que el hecho de que una canción sea de un artista puede hacer que determinados usuarios la escuchen y, por lo tanto, ser un factor clave para incluir en el modelo.

Otra conclusión respecto a la regresión lineal múltiple es que en todos los modelos la danzabilidad toma un valor positivo y elevado. Esto nos conduce a hacernos una idea de que cuanto más bailable sea una canción, más popularidad llegará a conseguir.

Análisis Discriminante:

Tras probar varios modelos de regresión, pensamos que la mejor opción sería realizar un Análisis Discriminante. La variable popularidad inicialmente es numérica por lo que si queremos realizar este método supervisado será necesario su recodificación en intervalos. Además, al igual que con los modelos de regresión hemos probado varios modelos y aprendido de los errores que cometíamos en el camino. Estos modelos han variado según las variables introducidas y según la partición en tramos de popularidad. Así pues, el primer modelo construido fue un modelo en el que partimos la variable independiente en tres intervalos aproximadamente iguales, es decir, "0-30", "30-70" y "70-100".



El resultado obtenido ha sido el mejor modelo de todos los que hemos creado, llegando a obtener un porcentaje de aciertos del 81% en los datos guardados para realizar la prueba y mostrando una *area under the curve (ROC)* de 0'79 mostrando unos buenos resultados.

Figura 10. Gráfica area under curve ROC para nuestro modelo discriminante

No obstante, nuestro modelo cuenta con un problema: las clases no se encuentran balanceadas (hay muchas canciones en el tramo 0-30 mientras que hay muy pocas en el tramo 70-100). Esto indica que el modelo tiende a predecir observaciones con popularidad baja y la potencia que tiene para predecir las altas es nula. Nuestro objetivo era predecir canciones con popularidad alta, por lo que este modelo no nos acabó resultando útil.

Debido a esto, realizamos nuevos modelos evitando este error, como por ejemplo creando dos clases, "0-10" y ">10". Los resultados obtenidos en este modelo resultaron tener una precisión más baja (0'6) y una *area under the curve* más baja (0'63), pero su potencia para predecir canciones con popularidad alta aumentaba, viéndose esto reflejado en la sensibilidad (0'56, mucho mayor que la sensibilidad del primer modelo desbalanceado, 0.00) y acercándonos así a nuestro objetivo.

Los resultados nos motivaron a seguir pensando en nuevos planteamientos que reduzcan los intervalos de popularidad elevada. Así pues, nuestra propuesta fue hacer un Análisis Discriminante dentro del Análisis Discriminante. En otras palabras, eliminamos de nuestro conjunto de datos todas las canciones con una popularidad inferior a 10 y nos quedamos con el resto. Sobre este conjunto recodificamos los datos teniendo en cuenta que estuvieran balanceadas las clases y, por tanto, partimos los datos en los intervalos "10-30" y ">30". En este caso, la *accuracy* seguía siendo similar (0'62), pero la potencia del modelo para predecir canciones con el valor de la variable respuesta elevadas disminuye considerablemente (0'09), haciendo este modelo no muy útil para alcanzar nuestros objetivos.

De la misma manera y como bien hemos explicado anteriormente, se han realizado varios modelos más considerando diferentes variables a introducir en el modelo y viendo que podíamos aprender de cada uno y que aprendíamos del proceso llevado a cabo.

Las conclusiones son que si nos tuviéramos que quedar con un modelo sería con el modelo de 4 variables que tiene una menor complejidad y se obtienen prácticamente unos resultados idénticos al modelo creado con todas las variables para el Análisis Discriminante con partición de la popularidad en tramos de “0-10” y “>10”. Además, sabremos que en el primer Análisis Discriminante que hagamos tendremos una mayor potencia para predecir datos con popularidad mayor que 10, pero que en el segundo modelo que generamos (con popularidad “10-30” y “>30”), la sensibilidad se reduce considerablemente y no nos ayuda a predecir datos con popularidad mayor que 30 de manera correcta. Así pues, proponemos estos modelos, teniendo en cuenta todo lo mencionado anteriormente y recalcamos que para mejorar en un futuro y crear un modelo mejor es necesario la introducción de nuevas variables como el artista o el año en que salió la canción, entre otras.

Se puede encontrar un análisis detallado con todos los modelos probados en el Anexo VIII.

4.3. Recomendador de canciones

Nuestra última meta era clasificar las canciones dependiendo de sus atributos musicales, y una vez realizado este análisis, ser capaces de recomendar canciones que tengan características muy similares, a partir de una que indique el usuario.

Para este objetivo hemos utilizado distintos métodos de clustering, un método estadístico no supervisado. Planteamos el método de Ward, para un clustering jerárquico, y por otra parte los métodos de k-means y PAM (k-medoides) en relación con los no-jerárquicos.

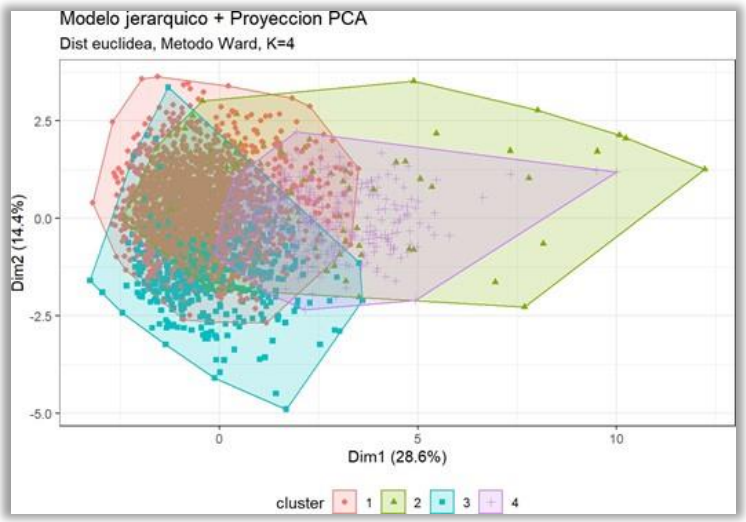
Inicialmente, decidimos calcular las distancias entre nuestras canciones para determinar si se podía observar cierta agrupación en los datos. Hemos decidido usar la distancia euclídea ya que nuestros datos no cuentan con características lo suficientemente complejas como para aplicar distancias como la de Manhattan o de Mahalanobis. El estadístico de Hopkins medio 0.8363, muy cercano a 1, con lo que se concluye que las canciones pueden agruparse en clusters.

En el primer intento filtramos las canciones para quedarnos con aquellas que tuvieran un valor de popularidad > 70, aunque más adelante ideamos un método de filtrado más complejo.

Figura 11. Matriz de distancias para las canciones con popularidad mayor que 70

4.3.1. Primera aproximación: método de Ward

Inicialmente se propuso un clustering jerárquico[3], aun sabiendo que este tipo de agrupamiento no es común en conjuntos de datos como el que tenemos (es decir, datos que no tienen atributos aglomerativos).



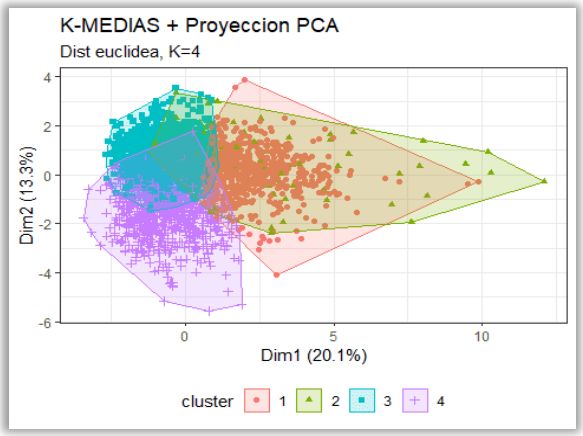
Como era de esperar, muchas canciones no quedaron bien agrupadas, existiendo un gran solapamiento entre clusters, con lo que, como habíamos previsto, los métodos jerárquicos no son adecuados para nuestro objetivo.

Figura 12. Proyección PCA de 4 clusters de un agrupamiento jerárquico mediante el método de Ward

4.3.2. Método de k-medias y k-medoides

Respecto a los métodos de partición, ambos métodos devuelven un número óptimo de clusters diferente: para k-medias el número es 4, mientras que para k-medoides el número es 2.

Con la misma cantidad de clusters que en el método de Ward, parece que k-medias hace la agrupación un poco mejor. No obstante, sigue existiendo un gran solapamiento: como se puede apreciar, la mayoría de las canciones caen en dos o más grupos diferentes, lo que dificulta su clasificación final.



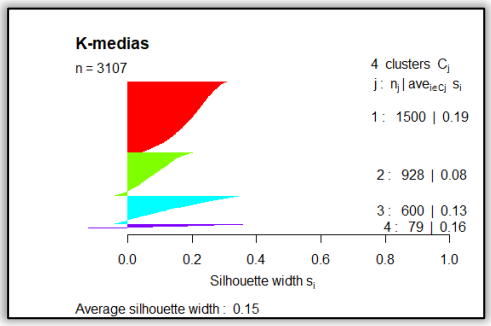


Figura 13. A la izquierda, protección PCA de 4 clusters de un agrupamiento de partición k-medias. A la derecha, el gráfico de silhouette para el mismo agrupamiento

[3] Un clustering jerárquico busca agrupar los datos en conjuntos que estén relacionados unos con otros mediante una jerarquía.

4.3.3. Clustering de clusters

Dado que no veíamos suficiente claridad en los grupos, a partir de los clusterings PAM y k-means, decidimos realizar agrupaciones dentro de los propios clusters.

Este es uno de los ejemplos. Para ser menos restrictivos, el clustering PAM original partía de sólo 2 grupos, y realizamos una nueva agrupación sobre el que más individuos tenía. Sorprendentemente, nos recomendaba elegir 5, 6 y 7 clusters, con resultados más bien deficientes. Era muy difícil agrupar las canciones en pocos grupos, por lo que nos planteamos mejorar el sistema de filtrado y elección de las variables, ya que puede que no todas fueran necesarias.

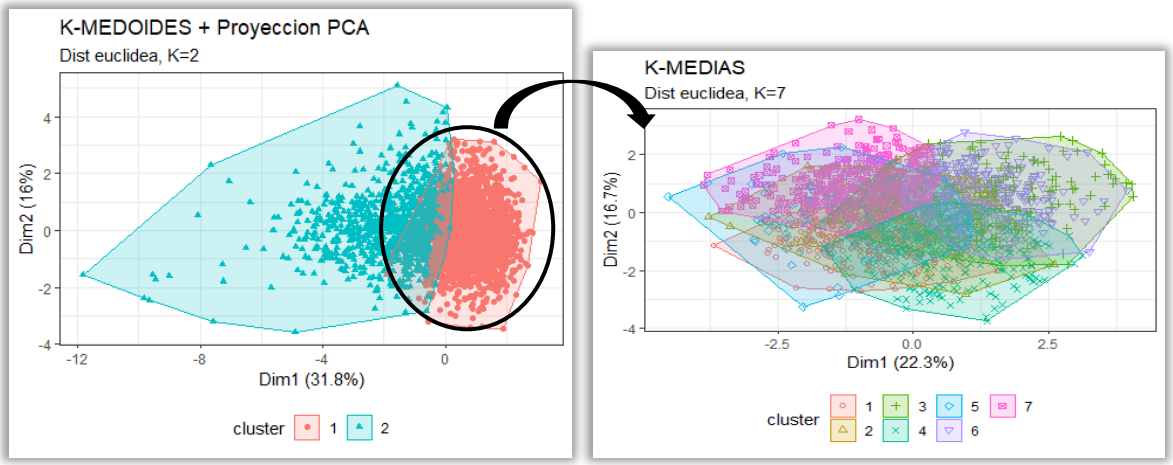


Figura 14. Proceso de clustering de clusters: partiendo de un agrupamiento de 2 clusters, aplicamos el mismo método, pero para las canciones que caen en el cluster 1, proporcionando ahora 7 grupos diferentes

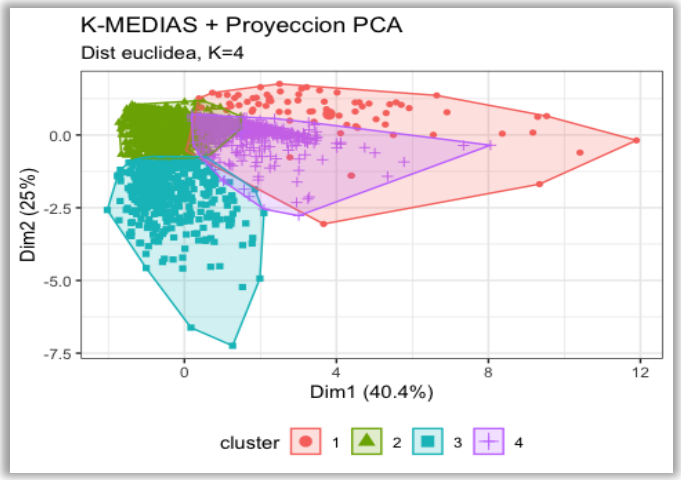
La agrupación, llegado este punto era decente, pero sobre ajustado a los datos ya que requería de muchas divisiones.

4.3.4. Clustering final: agrupamiento con variables incorrelacionadas

Con estos tres previos clusterings pudimos comprender lo siguiente: al existir canciones muy variadas en sus atributos musicales, existían demasiados factores para obtener un agrupamiento "limpio". Debido a ello, se tomaron dos decisiones:

1. Se añadirá un nuevo filtro para contar con un menor número de canciones. Se decidió filtrar por la duración de la canción, ya que algunas pistas eran podcasts de una hora o recopilaciones de muchas canciones, alterando los resultados.
2. Debido a la gran cantidad de variables, se escogerán aquellas que estén poco correlacionadas entre sí.

Por lo tanto, nuestro conjunto de datos final para hacer el clustering fueron aquellas canciones entre 30 segundos y 5 minutos, con las variables 'acústica', 'instrumentalidad', 'speechiness' y 'volumen'. El resultado del clustering es mucho más satisfactorio.



Como se puede observar, el solapamiento es mucho menor que en los anteriores, y la mayoría de las canciones caen en un solo cluster.

Vamos a ver ahora como este resultado nos ayuda a encontrar perfiles de canciones similares.

Figura 15. Proyección PCA para un agrupamiento k-medias con variables incorrelacionadas

Gracias a este clustering diferenciamos cuatro grupos que separan las canciones por los valores de sus variables, destacando en cada uno de los conjuntos unas características concretas.

Por ejemplo, el primer cluster destaca por su alto valor de instrumentalidad: es decir, incluye pistas de audio con poco contenido vocal y con bases musicales muy marcadas. En el segundo, las 4 características tienen valores similares y, por tanto, agrupamos en él las canciones que tienen tanto una base instrumental como vocal. En el tercero ocurre todo lo contrario al grupo 1: en este caso, la gran mayoría son pistas con fuerte contenido vocal. Es por esto que encontramos aquí grandes éxitos de rap y trap estadounidense o latino.

En el último cluster están representadas las pistas con números elevados en la variable que mide lo acústica que es la canción. Además, estas pistas tienden a tener valores bajos en 'loudness', lo que implica que en general son canciones poco ruidosas o estridentes.

	acousticness	instrumentalness	speechiness	loudness
1	0.6842331	5.7777741	-0.3187100	-1.346870413
2	-0.4733182	-0.1550215	-0.3237321	0.360017132
3	-0.1520269	-0.1737853	2.0942236	0.001440251
4	1.4428269	-0.1200889	-0.4333475	-0.917763687

Figura 16. Tabla con los valores de los centroides de los clusters en cada una de las variables usadas para el mismo

Esta división de los datos en cuatro grupos nos permite clasificar una nueva canción según su cercanía a cada uno de los centroides de estos. De esta forma, damos lugar a un sistema de recomendación de canciones en el que encontraremos la canción más próxima a la introducida, en un subconjunto de datos formado por las canciones del cluster al cual la nueva canción ha sido asignada.

Se puede encontrar cada método de clustering empleado y su análisis y evaluación en profundidad en los Anexos IX, X, XI y XII.

5. Conclusiones

Desde el principio del proyecto tuvimos claros los objetivos que íbamos a tener y para lograrlos tuvimos que hacer frente a distintas dificultades como hemos estado explicando a lo largo de esta memoria.

Primero de todo, en cuanto a los datos, hemos tenido una gran capacidad de adquisición gracias a las distintas bases de datos encontradas y nuestras habilidades a la hora de usar API REST. Obtener y estudiar los datos finales fue un reto a nivel computacional, ya que nunca habíamos trabajado con un volumen tan grande de datos.

Sin embargo, hemos de indicar que nuestras metas podrían haber abarcado ámbitos óptimos en resultados, como podría ser un recomendador personalizado para cada usuario, si también hubiésemos tenido la posibilidad de acceder a datos de usuarios, más personales.

Por otro lado, queremos destacar el enorme trabajo realizado con distintas herramientas tanto exploratorias como de análisis de los datos, como han sido la regresión, métodos discriminantes o clustering. La idea principal obtenida de estos estudios ha sido la dificultad de encontrar unos resultados que estuviesen a la altura de nuestras exigencias y objetivos principales.

Conclusiones de los objetivos

Gracias al PCA por décadas, hemos podido observar que ciertas características musicales persisten el paso del tiempo: las canciones siguen siendo altas, enérgicas, positivas y bailables. No obstante, a lo largo de los últimos años, algunos otros atributos han cambiado gratamente, siendo un ejemplo la aparición de canciones con contenido explícito o los cambios en la tendencia de la duración de estas. Se puede decir, en conclusión, que la diferencia está en los detalles.

El análisis discriminante confirma lo sospechado por el PCA por décadas, las canciones bailables, dado que son más comunes en reuniones sociales, conciertos y fiestas, y quizá por el creciente uso de aplicaciones como TikTok, tienen una ventaja para llegar a más gente, hacerse virales y, por tanto, ser populares. Aun así, los atributos musicales no son suficientes para modelar a la perfección la popularidad.

Tras muchos intentos hemos logrado un sistema de recomendación basado en las características musicales. Pero como hemos podido comprobar, simplemente las cualidades musicales quedan muy lejos de definir una canción: otras cosas como la propia letra de la canción pueden ser interesantes para recomendar pistas del mismo género e idioma. Se puede encontrar el código de este recomendador, así como una explicación de este, en el Anexo XIII.

6. Lecciones aprendidas para mis futuros proyectos de ciencia de datos

En el mundo de la música, puede que 1 millón de canciones no sean una gran cantidad, pero a la hora de analizarlos y procesarlos ha sido prácticamente imposible tenerlos en cuenta todos a la vez. La dificultad residía en los métodos estadísticos utilizados para lograr los objetivos: RStudio sí que nos permite el procesamiento de tantos datos al mismo tiempo, pero la capacidad computacional de nuestros ordenadores quedaba muy lejos de la necesaria.

Por lo tanto, sería interesante volver a realizar un proyecto similar a este dentro de un tiempo, cuando podamos aplicar algunos métodos quizás más aptos para nuestro volumen de datos y los objetivos del proyecto, y, sobre todo, dispongamos de la capacidad de procesamiento que requieren estos data sets.

Además, hemos aprendido a no ver los métodos que aplicamos por separado, sino que muchas veces es útil combinarlos de forma que simplifiquen los análisis para obtener resultados más claros. Por ejemplo, en nuestro caso hemos realizado clusterings sobre resultados de un clustering.

Por otra parte, en cuanto al software utilizado, destacamos el trabajo con Python para extraer datos de las API (en este caso Spotify API) con la finalidad de actualizar nuestros datos de popularidad y corregir valores faltantes. Este paso ha sido primordial para poder extraer resultados actuales y nos ha hecho darnos cuenta de lo importante que es saber el origen de nuestros datos e intentar corregir los posibles errores o desactualizaciones en estos, ya que, si los datos están desactualizados, sea cual sea el proyecto, todos los resultados que obtengamos no van a útiles ni van a despertar ningún interés.

Con todo lo expuesto anteriormente, concluimos en que a pesar de que nos hubiera gustado obtener unos resultados mucho más sofisticados (incluso trabajando con datos de usuarios de Spotify los cuales no hemos podido obtener por temas de privacidad), este proyecto nos ha servido para coger experiencia trabajando con datos reales, donde hemos tenido que realizar muchas operaciones de limpieza y actualización de datos, así como dificultades a la hora de procesar muchos análisis. Por eso, valoramos mucho esta experiencia que hemos tenido, ya que hasta ahora casi siempre habíamos trabajado los métodos para el análisis con conjuntos de datos ya procesados o preparados para obtener algún resultado concreto en las prácticas de otras asignaturas.