Grupo 10

TPE - Fundamentos de la Ciencia de Datos

Análisis Exploratorio de Canciones de los Años 70: Patrones, Tendencias y Conclusiones

Participantes:

- Etcheverria, Victoriano.
- Rapela, Guadalupe.
- Roumec, Iñaki.

ÍNDICE

Introducción

La música de los 70 se caracterizó por su diversidad y riqueza. Géneros como el *rock*, el *funk*, el *soul*, el disco y el *folk* se consolidaron como expresiones populares, con artistas que exploran nuevos sonidos y letras que reflejaban la realidad social y cultural del momento.

Con la música de los 70 la gente bailaba al ritmo del *rock*, *soul*, *rock country*, *punk*, incluso de las nuevas propuestas musicales de *jazz* de la época. Ésta fue la época de la discoteca, sitio nocturno donde adquiere vida el género disco.

A través de este informe, se guiará al lector en un análisis extenso de una parte de lo que fue la música de los 70, analizando distintos factores y llevando el foco hacia qué era lo que motivaba a las personas a bailar al compás de la canción.

Comprensión del Dominio.

Previo al análisis de los datos, es necesario entender conceptos básicos del dominio. A continuación, se brinda una breve explicación de las variables más abstractas y propias de la música abarcadas en el *dataset*.

La tonalidad o *key* se refiere al conjunto de notas y acordes que conforman una pieza musical, basándose en una nota principal llamada *tónica*. La teoría musical occidental distingue 12 tonalidades:

1. C (Do)
2. C# (Do#)
3. D (Re)
4. D# (Re#)
5. E (Mi)
6. F (Fa)
7. F# (Fa#)
8. G (Sol)
9. G# (Sol#)
10. A (La)
11. A# (La#)
12. B (Si)

A la izquierda de la enumeración, puede observarse el **cifrado americano** o **cifrado anglosajón** que utiliza las primeras siete letras del alfabeto para nombrar a las notas o acordes. Es una nomenclatura más simple y utilizada que la que puede verse entre paréntesis a la derecha, denominada *nomenclatura convencional*. También, puede notarse al lado de algunas tonalidades el símbolo de **sostenido** (#) que, sin entrar en mucho detalle, indica una diferencia de un semitono o medio tono. Esto señala que, por ejemplo, Do# es un poco más alto que Do.

A su vez, las tonalidades pueden expresarse en modo o tono (*mode*, en inglés) mayor o menor, lo cual dota a la música de distintas características emocionales. En general, las tonalidades mayores suelen ser más alegres y optimistas, mientras que las menores tienden a ser más oscuras o melancólicas.

El pulso es la unidad básica que se emplea para medir el tiempo, es el latido de cada pieza. A la velocidad o rapidez del pulso se denomina tempo. Esta es otra característica importante que define la atmósfera y emoción de una pieza musical. Se mide en pulsaciones por minuto (*BPM*, por sus siglas en inglés: *beats per minute*), que indica cuántos pulsos hay en un minuto. Por ejemplo, un tempo de 60 BPM significa que hay un pulso por segundo, mientras que 120 BPM indica dos pulsos por segundo. Un tempo rápido es asociado a piezas musicales más enérgicas y vibrantes, mientras que un tempo lento a piezas más reflexivas o melancólicas.

A la hora de hacer música lo que se hace es crear patrones con los pulsos. Es decir, se acentúan o marcan con más intensidad uno o varios pulsos. A cada uno de los grupos formados por el pulso acentuado y los que vienen después se les llama **compás**.

La firma de tiempo o *time signature* está compuesta de dos números: el numerador, el cual indica cuántos pulsos hay en cada compás; y el denominador, el cual indica qué tipo de nota equivale a un pulso.

Clave de sol con una firma de tiempo 4/4 en un pentagrama.

Las conceptos definidos anteriormente, pueden asociarse con las variables halladas en el *dataset* de la siguiente forma:

- Lo que el *dataset* define como *time_signature* es el compás.
- Lo que el *dataset* define como *key* es la tonalidad.
- Lo que el *dataset* define como *mode* es el tono.
- Lo que el *dataset* define como *tempo* es el tempo.

Otras variables importantes de entender en este contexto y que se hallan en el *dataset* son:

- *Track*: texto libre que representa el nombre de la pista.
- *Artist*: texto libre que representa el nombre del artista.
- *Duration*: texto en formato horario que representa la duración de la pista en minutos.
- *Danceability*: variable cuantitativa continua que representa qué tan bailable es una canción, basada en el *tempo*, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general.
- *Energy*: variable cuantitativa continua que representa una medida de intensidad y actividad en la canción, donde los valores más altos indican una pista más enérgica.
- *Loudness*: variable cuantitativa continua que representa el volumen promedio de la canción, medido en decibelios (dB).
- *Speechiness*: variable cuantitativa continua que representa la presencia de palabras habladas en una pista. Valores más altos indican cualidades más parecidas al habla.
- *Acousticness*: variable cuantitativa continua que representa una medida de la calidad acústica de la pista. Valores más altos indican una mayor probabilidad de ser acústica.

- *Instrumentalness*: variable cuantitativa continua que la presencia de voces. Valores más altos representan pistas más instrumentales.
- *Liveness*: variable cuantitativa continua que representa una medida de la probabilidad de que la pista se haya interpretado en vivo. Valores más altos indican más ruido de audiencia.
- *Valence*: variable cuantitativa continua utilizada como medida de la positividad musical de la pista. Valores más altos indican música más positiva o alegre.
- *Popularity*: variable cuantitativa discreta que representa una puntuación que refleja la popularidad de la pista. Generalmente, basada en los recuentos de transmisiones y otras métricas.
- *Year*: variable cualitativa ordinal que representa el año en que se lanzó la canción.

Preparación de los Datos

El *dataset* brindado contiene un total de 980 registros y 17 columnas, donde cada columna corresponde a un valor asociado a cada una de las variables y conceptos mencionados anteriormente.

Adicionalmente, se utilizó un *dataset*, como punto de apoyo de análisis posteriores, se utilizó un *dataset* adicional con datos obtenidos de la plataforma *Spotify*, el cual abarca información de diversas pistas lanzadas entre los años 1920 y 2020 en distintos países e idiomas.

Limpieza del *Dataset*.

Previo al análisis de sus datos, se realizó su limpieza. Durante su transcurso, se identificó lo siguiente:

- **Pistas duplicadas en las que la única diferencia era el año de lanzamiento**. Se validó cuál es el año de lanzamiento correcto de la pista y se descartaron los registros incorrectos.
- **Canciones con todos sus datos iguales, exceptuando el artista**. Se verificó que ambos artistas contienen una canción con dicho nombre. Sin embargo, los demás datos no correspondían a ninguna de las dos pistas.
- **Títulos en los que se observa el nombre de dos canciones**. Al realizar una búsqueda sobre las canciones, notamos que el valor de la variable Duration concuerda con una sola de las duraciones o con ninguna de las dos, por lo que se decidió descartar estos registros.
- **Filas en la que la columna *Track* y *Artist* se detectaron incorrectamente**. En todas ellas, el texto que figura en la variable *Artist* es en realidad parte del nombre de la pista. La separación está realizada a partir de la palabra *by* (utilizada, en inglés, para especificar que algo fue realizado por alguien), la cual conforma también el nombre de la pista. Con esto en cuenta, se halló el artista de cada pista y se realizó la corrección de la información:

Preprocesamiento.

La variable *Duration* estaba originalmente en formato de texto, representada con valores como "3:45", donde "3" indica minutos y "45" segundos. Para facilitar su análisis, se decidió transformarla a un tipo de dato cuantitativo discreto, cuyo valor numérico indica la duración total de la pista en segundos.

Análisis Exploratorio de los Datos (EDA).

Habiendo realizado la limpieza de los datos, a partir del uso de herramientas estadísticas y de visualización, se realizó un análisis exploratorio de estos. El mismo fue dividido en tres fases:

1. Análisis univariado.
2. Análisis bivariado.
3. Análisis multivariado.

Se busca, así, tener una idea completa de cómo son los datos y cómo se relacionan entre sí para sacar conclusiones, construir hipótesis y, en apartados posteriores, poder validarlas.

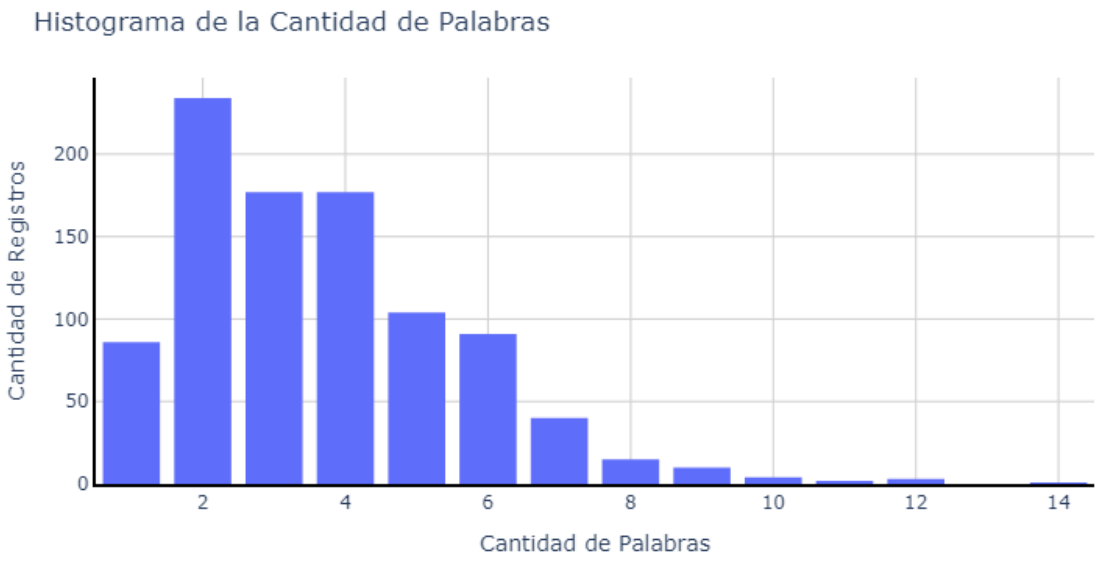
Análisis Univariado.

El análisis univariado consiste en el análisis individual y aislado de cada una de las variables.

A continuación, se presentan los resultados más relevantes hallados en el análisis:

Track.

- Se sospechó que la longitud del nombre de la pista podía llegar a tener relación con alguna de las demás variables. Por consiguiente, se añadió al conjunto de datos, para cada registro, una variable que indicará el número de palabras en la canción.
- Se determinó que la mayoría de las pistas contienen cinco o menos palabras en su título.



Histograma de la cantidad de palabras de un Track

- El nombre de todas las pistas del *dataset* se halla en inglés, exceptuando una: *Eres tú de Mocedades*, la cual fue muy popular en el público anglosajón.
- Se consideró interesante analizar las palabras más frecuentes en el título, haciendo a un lado preposiciones, artículos y otro tipo de palabras que no se consideraron relevantes. Para ello, se realizó un *WordCloud*:

WordCloud de la variable Track

- Se detectó una presencia elevada de palabras de índole romántico en los títulos, tales como *Love*, *Sweet*, *Girl*, *Baby*, etc.

- Se observó la presencia de modismos estadounidenses en los títulos, tales como *ain't* y *gonna*, lo que levantó la duda: *¿habrá una elevada proporción de artistas estadounidenses en el dataset?*

Artist.

- Utilizando la API de MusicBrainz, una enciclopedia libre que se encarga de la recolección de datos y metadatos acerca de pistas y sus artistas, se obtuvo la nacionalidad de cada artista. Con la información obtenida, fue posible validar que, efectivamente, la mayor parte de los artistas en el *dataset* son estadounidenses.

Distribución de los artistas por nacionalidad

Duration.

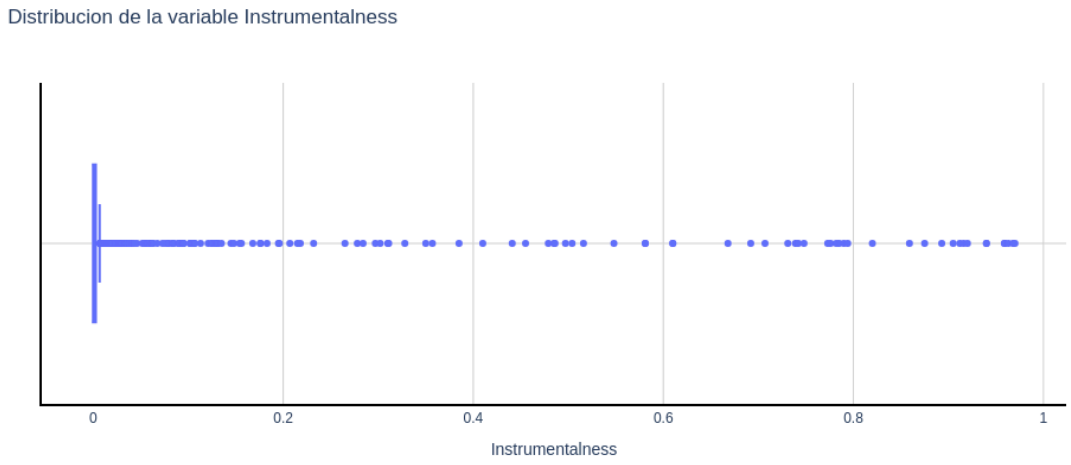
- Se identificaron pistas con duraciones considerablemente extensas, siendo una de hasta 26 minutos. Luego de analizarlas, se encontró que varias de ellas no eran en realidad pistas, sino álbumes, o su duración estaba erróneamente cargada. Por consiguiente, se descartaron del conjunto de datos.

Time Signature.

- Se identificaron canciones con compases de 1 y 5. Al realizar la búsqueda, se halló que dichos valores de compás son extremadamente raros y que no corresponden con las pistas en las que fueron hallados. Por consiguiente, fueron descartadas.
- El resto de las pistas tiene un compás de 3 o 4, lo que convierte a la variable en dicotómica.
- El compás predominante es el compás de 4, presente en casi un 95 % de las observaciones.

Instrumentalness

- Se halló una distribución particularmente extraña en la variable:



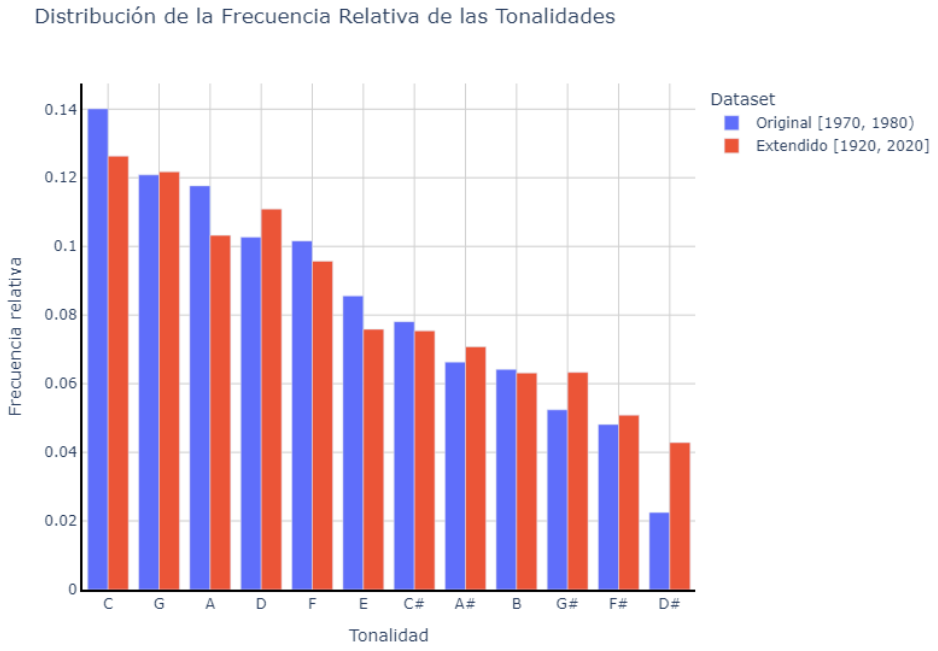
Distribución de la variable Instrumentalness

- Al escuchar pistas con valores de *Instrumentalness* altos y compararlas con aquellas con valores más bajos, no se halló ninguna diferencia en cuanto a los instrumentos, la presencia de voces ni ninguna otra característica. Se concluyó entonces que la variable no determina lo que dice determinar ni representa lo que su nombre indica.
- Debido a que se desconoce qué es lo que la variable describe y que, por lo tanto, no es posible comprobar su validez ni sacar conclusiones acerca de ella, fue descartada del análisis.

Key

- Se analizó la frecuencia de aparición de las distintas tonalidades, hallándose que C (Do, en nomenclatura convencional) es la tonalidad más frecuente, mientras que D# (Re sostenido), la más infrecuente.

- Con la finalidad de determinar si la frecuencia relativa de uso de las tonalidades en los 70 es representativa de los últimos cien años, se comparó la frecuencia relativa del *dataset* original con el extendido:



Distribución de la frecuencia relativa de las tonalidades.

- Es posible observar que la frecuencia relativa de las tonalidades del *dataset* es significativamente representativa de la frecuencia relativa de las tonalidades en los últimos cien años. La tonalidad D# es la única que pareciera no seguir este comportamiento.

Popularity

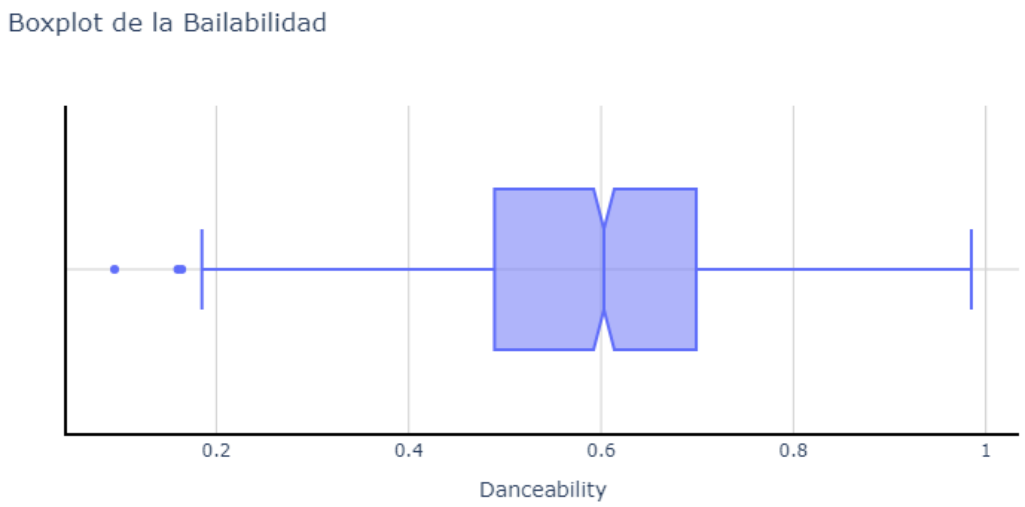
- Si bien constituye la variable de más utilidad en la industria musical, también es la más enigmática. Se observaron valores de popularidad extraños (canciones muy populares, con millones de reproducciones en distintas plataformas, con valores de popularidad considerablemente más bajos que canciones que no parecen ser conocidas). Debido a estas peculiaridades, la variable, quien había sido en un principio planteada como variable objetivo y a

partir de la cual se realizaría el análisis, fue descartada. Esto debido a que, a diferencia de las demás variables que se mantienen en el tiempo y la región, la popularidad no es así y, al no contar con la población ni momento de la que fue extraída, las conclusiones que se obtengan de ella no serían relevantes.

- Se detectaron registros con valores atípicamente bajos de la variable, los cuales fueron descartados al no ser posible comprobar su veracidad.

Danceability.

- Descartada la popularidad como variable objetivo, se decidió adoptar la bailabilidad en su lugar. Si bien no es tan importante en la industria como la popularidad, puede ser relevante si lo que se quiere, por ejemplo, es crear una pista con el objetivo de que sea bailada en los boliches. Además, la bailabilidad de algo no se halla tan fuertemente ligado a la región, a una población en particular o a situaciones sociales; lo que suele ser bailable en una parte del mundo o en una época también es bailable en otra.
- La variable toma valores entre 0 y 1, donde valores más cercanos a 1 indican mayor bailabilidad en la pista.
- La variable es un dato derivado del *tempo*, lo cual es importante considerar al momento de analizar relaciones de ella con otras variables.



Boxplot de la Bailabilidad

- Se detectaron valores atípicos en la variable que, posterior a su análisis, se concluyeron como verídicos.

Mode.

- Cerca del 75 % de las pistas del *dataset* se hallan en tono mayor.

Otras Variables Musicales.

(energy, speechiness, acousticness, liveness, valence, tempo, loudness)

- Se realizaron análisis gráficos y se calcularon medidas de asimetría para estudiar la distribución de estas variables. En general, las distribuciones de estas se hayan bastante sesgadas.
- Se identificaron valores atípicos en algunas de ellas. No obstante, al analizarlos de manera detallada, se concluyó que se trataban de datos verídicos.

Análisis Bivariado.

Como primer paso del análisis, se realizó un *heatmap*, utilizando la correlación de Pearson, para detectar correlaciones lineales entre las variables. Este análisis no fue de especial utilidad, ya que al observar el gráfico no se hallaron valores los cuales superaran el mínimo empírico de 0.7.

Aun así, se tomó la decisión de analizar los tres pares de variables con correlaciones más altas. Estos son:

- *Energy* y *Loudness*: con un valor de 0.66.

- *Acousticness* y *Energy*: con un valor de -0.59.

- *Danceability* y *Valence*: con un valor de 0.52.

El resultado del análisis determinó que ninguno de los pares de variables anteriores presenta una relación lineal lo suficientemente definida. La mayoría de ellas parecían ser meras nubes amorfas.

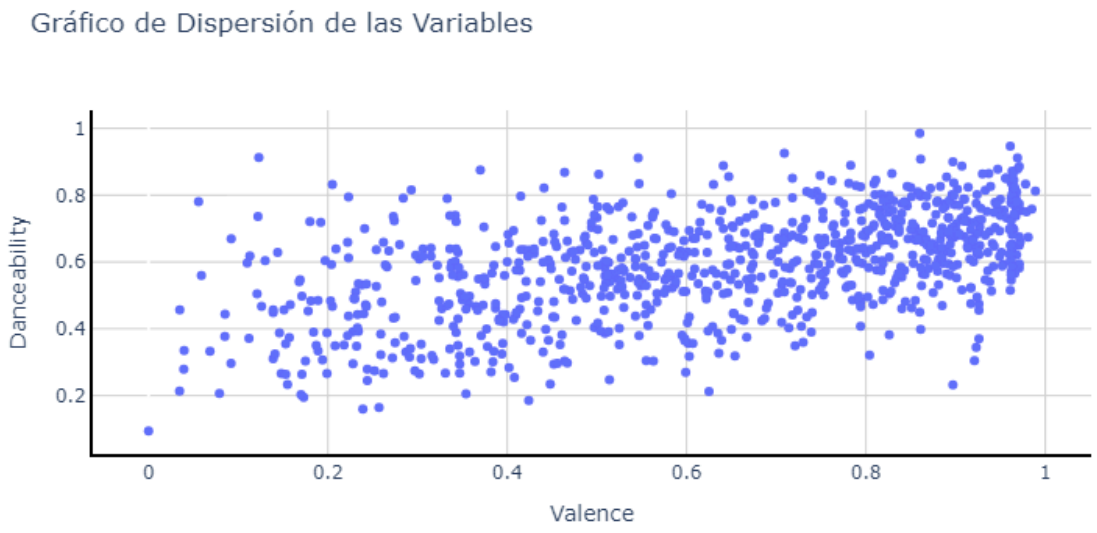
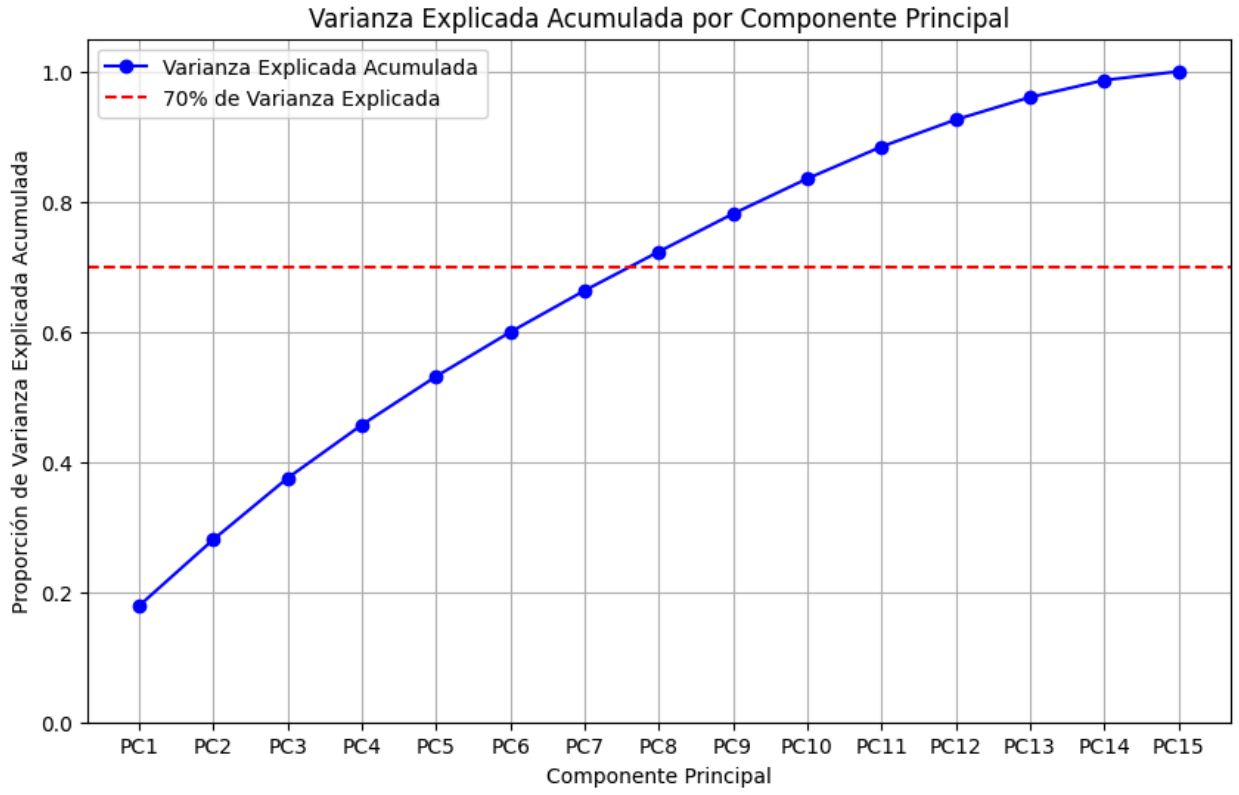


Gráfico de dispersión de las variables Danceability y Valence

Se analizó la posibilidad de la existencia de una relación no lineal entre las variables. Para ello, se analizaron detenidamente los gráficos de dispersión de cada posible par de variables. Si bien fue posible observar ciertos patrones entre algunas de ellas, la dispersión parecía corresponder más a una nube amorfa que a una relación claramente definida.

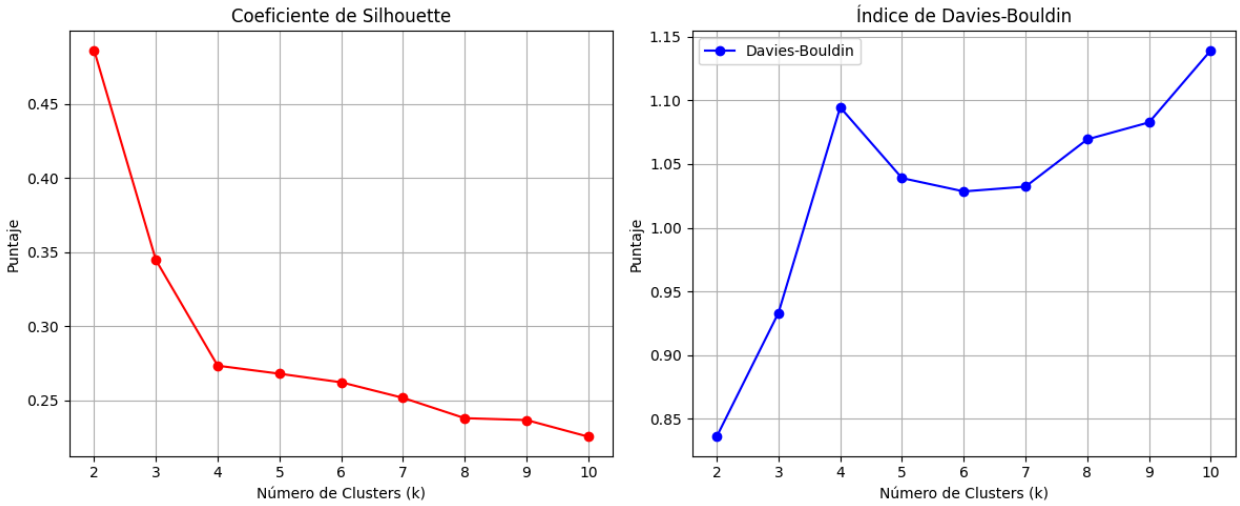
Análisis Multivariado.

Bajo la motivación de una mejor visualización y la búsqueda de agrupaciones, se evaluó la posibilidad de aplicar una reducción de la dimensionalidad en el conjunto de datos. Entre las opciones evaluadas, se halla PCA, la cual fue descartada puesto a que el número necesario de componentes para explicar una variabilidad significativa en los datos era muy alto y no podría ser visualizado gráficamente.



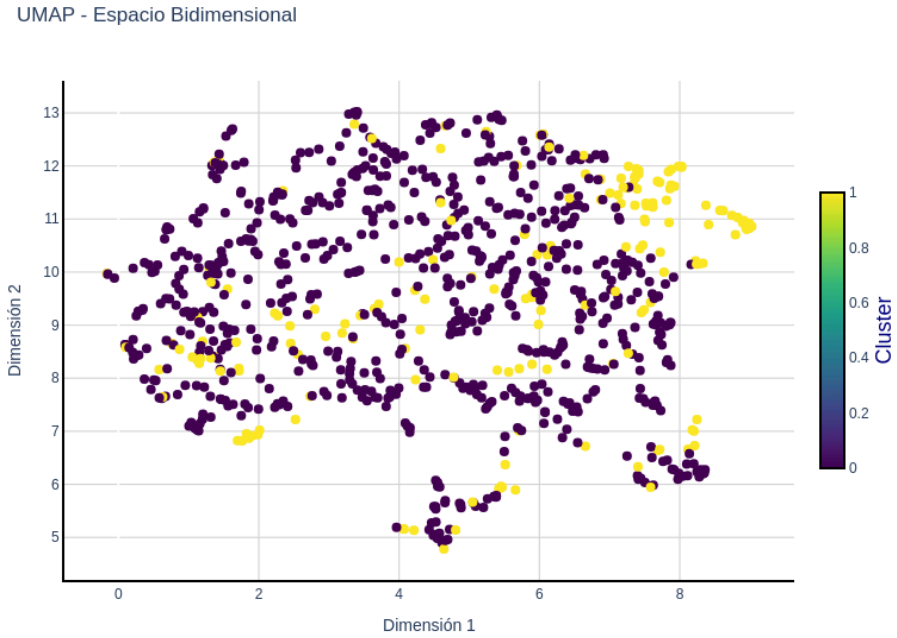
Varianza explicada acumulada por componente principal

Posteriormente, se optó por UMAP, una técnica de reducción de la dimensionalidad no lineal. Al realizar la reducción a un espacio bidimensional, no fue posible observar agrupaciones bien diferenciadas. Por consiguiente, en busca de estas, se realizó la aplicación del método de *clustering KMean*, seleccionando el número óptimo de *clusters* con ayudas gráficas como el *elbow plot* y utilizando métricas, tales como el coeficiente de Silhouette y el índice de Davies-Boulding.



Métricas que indican él número óptimo de clusters

No obstante, ni el *clustering* ni la reducción de la dimensionalidad permitieron la identificación de agrupaciones.



UMAP coloreado por clusters.

Adicionalmente, se indagó con otras técnicas y métodos, como la utilización de *hierarchical clustering* en lugar de *Kmean*, o la aplicación de UMAP a grupos específicos del conjunto de datos. No obstante, no fue posible obtener ningún resultado concluyente con ninguno de ellos.

Test de Hipótesis

A partir de lo recopilado en el análisis, se plantearon seis hipótesis para el *dataset* de estudio.

La **primera hipótesis** retoma lo mencionado en el *Análisis Univariado* y busca demostrar que la frecuencia relativa de las tonalidades en los 70 es representativa de su frecuencia relativa en los últimos cien años. Durante el análisis de la variable *Key*, utilizando herramientas de visualización, se verificó que así es para todas las tonalidades, exceptuando para D# (Re sostenido, de acuerdo al cifrado convencional), cuya representación no es clara de forma visual. Por lo tanto, la hipótesis planteada y no validada es: *El uso de la tonalidad D# (Re sostenido) en la década de los 70 no es representativo de su uso en los últimos cien años*.

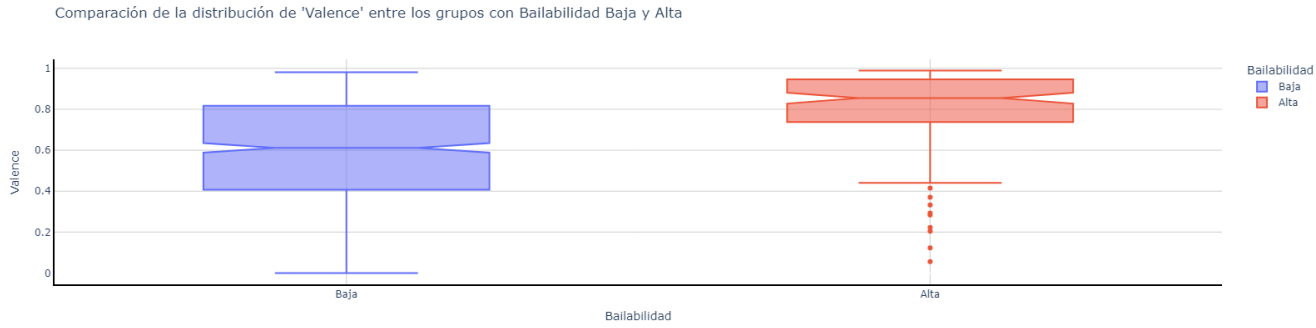
Las cinco hipótesis restantes, las cuales fueron validadas, se tratan en los siguientes apartados y giran en torno a la variable objetivo: la bailabilidad.

Aclaración: en todos los *tests* realizados, el nivel de confianza utilizado es del 95 %.

Hipótesis 2: La alegría del baile.

Una canción más alegre (con valores más altos de Valence) es más bailable.

Esta hipótesis surge de manera intuitiva, ya que parecería lógico pensar que para que una canción genere ganas de bailar, debiera transmitir alegría. Un buen ejemplo de esto es "Get Down, Get Down" de Joe Simon, una de las canciones con mayor bailabilidad del *dataset*, que al escucharla inspira alegría y movimiento, la cual tiene una valencia de 0.86. En contraste, "Ain't No Sunshine" de Bill Withers, es una canción que, con un valor de Valence de 0.00001, está muy lejos de motivar a bailar.



Comparación de la distribución de "Valence" entre grupos de bailabilidad baja y alta.

Se observó que las canciones con alta bailabilidad tienden a tener valores más altos de *Valence*, y que esta diferencia es significativa. Esto se confirma al analizar los *notches* de cada boxplot, los cuales se hallan considerablemente diferenciados, sugiriendo una notable diferencia entre los grupos.

Para respaldar esta observación, se realizó un test de hipótesis dividiendo las canciones en dos grupos acorde a un criterio arbitrario: aquellas con alta bailabilidad (*Danceability* ≥ 0.75) y las de baja bailabilidad (*Danceability* < 0.75). Tras realizar los *tests* de normalidad y homocedasticidad para determinar el tipo de *test* a utilizar, se aplicó el *test* de Kruskal-Wallis. Este análisis produjo un *p-valor* cercano a 0, lo que conlleva rechazar la hipótesis nula. Por lo que se concluye que **la bailabilidad de una canción está relacionada con sus valores de *Valence***, siendo significativamente mayor en canciones de alta bailabilidad.

Hipótesis 3: Los pies se cansan.

Las canciones de gran duración no son adecuadas para bailar.

Esta hipótesis surge de la reflexión sobre las características que hacen que una canción sea o no bailable. Se considera que la duración podría ser un factor determinante, ya que una pista extensa podría resultar monótona y provocar que el oyente se canse al pasar de los minutos. Así, se propuso que las canciones de larga duración serían menos bailables que las de duración más moderada.

Para poner a prueba esta hipótesis, se utilizaron las distinciones realizadas en la hipótesis anterior, en la que se diferencian las canciones en dos grupos, de acuerdo a si su bailabilidad es alta o baja.

Primero, se evaluó la normalidad de los datos con el test de Shapiro-Wilk, el cual arrojó un p-valor cercano a 0. Esto indica que la duración de las canciones no sigue una distribución normal. Posteriormente, evaluamos la homocedasticidad (igualdad de varianzas) usando el *test* de Levene, cuyo p-valor fue superior al 5%, permitiéndonos aceptar la hipótesis, por lo que las variables presentan homocedasticidad.

Teniendo en cuenta los resultados anteriores, se eligió utilizar el *test* de Mann-Whitney U para analizar la hipótesis propuesta. Este test se aplica cuando no hay normalidad en los datos, y su hipótesis nula establece que no existen diferencias significativas entre los dos grupos en cuanto a la duración de las canciones. Tras realizar el cálculo, el test de Mann-Whitney arrojó un p-valor de 0.318, lo que indica que no existe suficiente evidencia para rechazar la hipótesis nula.

Como conclusión, no se halló una diferencia significativa en la duración entre canciones bailables y no bailables, lo cual lleva a rechazar la hipótesis planteada. Esto sugiere que, al menos en los datos analizados, **la duración de una canción no parece ser un factor determinante en su bailabilidad**.

Un posible motivo para esta falta de relación es que, aunque una canción sea larga, puede contar con un estribillo o una sección rítmica repetitiva que sea altamente bailable. Estos elementos permiten mantener el interés y la energía en la pista de baile, independientemente de la duración total de la pista. Así, una canción extensa podría igualmente funcionar como pieza bailable.

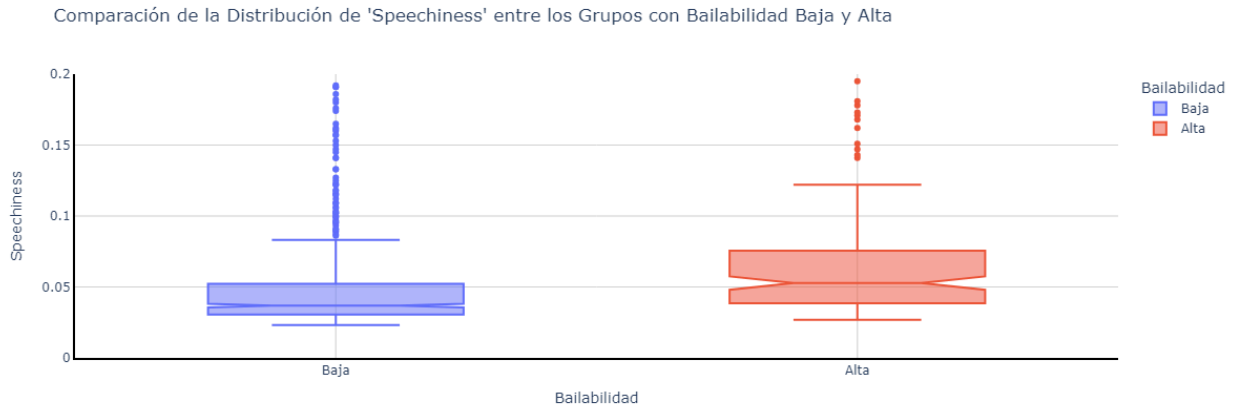
Hipótesis 4: La importancia del ritmo.

Las canciones con menor nivel de habla son más bailables.

Esta hipótesis surge de la idea de que una canción, para ser bailable, debería contener menos contenido hablado, considerando que géneros donde las canciones son muy habladas, como el *rap*, tienden a no serlo.

Para probar esta hipótesis, se evaluó la normalidad de los datos mediante el *test* de Shapiro-Wilk, el cual arrojó un p-valor cercano a 0 en ambos casos. Esto permitió rechazar la hipótesis nula de normalidad. Posteriormente, se realizó una prueba de homocedasticidad utilizando el *test* de Levene, que dio como resultado un p-valor de 0.08, indicando que ambos grupos presentan una varianza similar y cumplen con el supuesto de homocedasticidad. Dadas estas condiciones, se recurrió al *test* de Mann-Whitney U para comparar el nivel de *speechiness* entre ambos grupos.

El test de Mann-Whitney U, cuya hipótesis nula establece que los centros de las distribuciones no difieren, arrojó un p-valor cercano a 0. Esto permitió rechazar la hipótesis nula, indicando una diferencia significativa en el nivel de *speechiness* entre las canciones bailables y no bailables.



Comparación de la distribución de "Speechiness" entre los grupos con bailabilidad baja y alta.

Al analizar un *boxplot* comparativo de los grupos, se observa que, efectivamente, existe una diferencia significativa. No obstante, es posible notar que, en las canciones con mayor bailabilidad, los valores de *speechiness* son, en general, más altos, lo que sugiere que el nivel de *speechiness* influye en la bailabilidad, pero de una forma inversa a la planteada en la hipótesis inicial, dado que **las canciones más bailables resultan ser, en realidad, más habladas**.

Hipótesis 5: No me bailes con ese tono.

Las canciones compuestas en tono mayor (mode=1) son más bailables que las canciones en tono menor (mode=0).

Esta hipótesis surge de la lógica y de una observación inicial de algunos registros del *dataset*. Al escuchar ejemplos de ambos tipos de canciones, es razonable suponer que las canciones en tono mayor son más bailables que las de tono menor. Por ejemplo, al comparar "The Guess Who - No Time", la cual posee tono mayor, y "Rainy Night in Georgia", de tono menor, se percibe que la primera es mucho más bailable. Además, es común escuchar que el tono mayor se asocie a pistas más alegres y enérgicas, lo que podría resultar en más ganas de bailarla.

Para evaluar la hipótesis, primero, se evaluó si los grupos de canciones en tono mayor y tono menor son apareados o desapareados. Como las canciones en tono mayor y tono menor pertenecen a conjuntos sin individuos en común, se trata de grupos desapareados.

Primero, se verificó la normalidad en ambos grupos usando la prueba de Shapiro-Wilk. Con un p-valor de 0.001 para mode=1 y un p-valor casi nulo para mode=0, ambas pruebas indicaron que los grupos no siguen una distribución normal.

Luego, se evaluó la homocedasticidad usando la prueba de Levene, cuya hipótesis nula afirma que los grupos presentan varianzas iguales. Dado que el p-valor fue 0,64 (mayor a 5%), se valida el supuesto de homocedasticidad.

Finalmente, se realizó el *test* de Mann-Whitney U, cuya hipótesis nula propone que no hay diferencia en la tendencia central de los dos grupos. Con un p-valor igual a 0,024 (menor al 5%), se rechaza la hipótesis nula y se concluye que, efectivamente, **el nivel de bailabilidad de las canciones difiere según el tono (mayor o menor) de la canción**. Pero, antes de finalizar, se debe comprobar que la diferencia significativa sea para él lado planteado en la hipótesis, es decir, que las canciones en tono mayor tengan valores más altos de bailabilidad, para esto se agregó él parámetro alternative = 'less' que quiere decir que él primer grupo (moda 0) tiene muestras con menos bailabilidad que él segundo (moda 1), realizando de vuelta él test, lanzó un p valor muy cercano a 1, expresando que él grupo que tiene más bailabilidad, no es él que tiene moda 1, sino él que tiene moda 0.

Se puede concluir que si hay una diferencia de bailabilidad si la canción está hecha en un tono menor o mayor y que al contrario de lo que se pensaba, **las canciones más bailables son las compuestas en un tono menor.**

Hipótesis 6: ¿Qué es esto? ¿Se puede bailar?

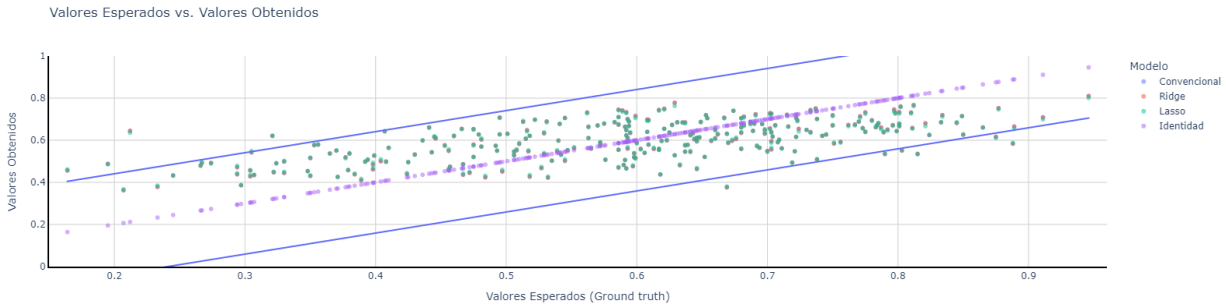
Dadas los resultados de las hipótesis anteriores, la variable Danceability debería poder ser predicha linealmente con éxito por el conjunto de variables del dataset, ***siendo las variables Valence, Mode y Speechiness las más aportativas a la predicción****.*

Para comprobar esta hipótesis, se utilizaron tres modelos de regresión lineal distintos: el modelo convencional, Ridge y LASSO. Inicialmente, se realizó una partición de los datos en conjuntos de prueba y entrenamiento, asignando el 30% de los datos al conjunto de prueba. Del conjunto de entrenamiento, se generó un subconjunto adicional de validación que comprende el 10% de sus datos. Posteriormente, se estandarizaron los datos y se determinó el valor óptimo de *alpha* para los modelos Ridge y LASSO. Tras el ajuste del hiperparámetro, se entrenaron los tres modelos y se compararon sus métricas.

	MAE en test	MSE en test	RMSE en test
Modelo convencional	0.10125	0.01512	0.12297
Ridge (mejor alpha)	0.10125	0.01512	0.12297
Lasso (mejor alpha)	: 0.10144	0.01517	0.12318

Comparación de métricas de evaluación.

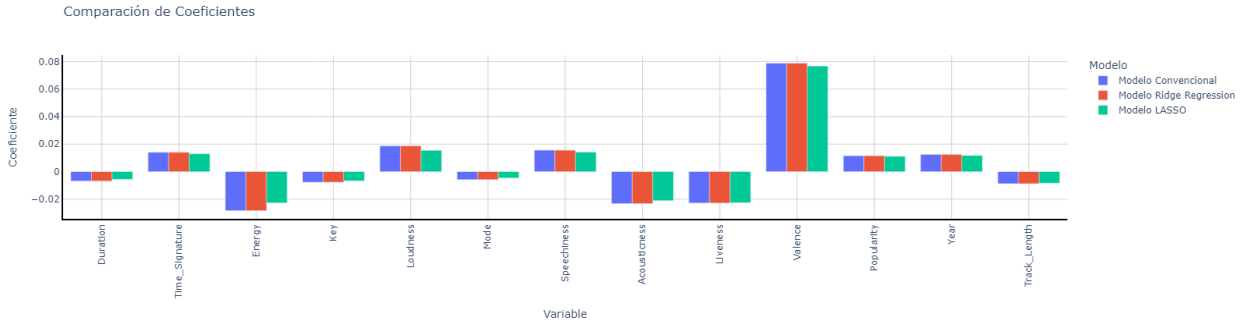
Los resultados muestran que las métricas de los modelos convencional y Ridge son prácticamente idénticas, lo que indica que ambos presentan un rendimiento similar en la predicción de la variable bailabilidad. En contraste, el modelo LASSO obtuvo métricas más elevadas, sugiriendo que es el menos adecuado entre los tres modelos. Aplicando el principio de parsimonia, se optó por el modelo convencional, debido a su simplicidad con respecto a Ridge, puesto a que este último implementa un regularizador.



Para evaluar qué tan bien predice el modelo, se realizó un gráfico en el que se se compararon los valores predichos con los esperados. Los puntos denominados *Identidad* indican lo ideal: que el modelo haya predicho el valor exacto de bailabilidad. La *identidad* está formada por las muestras del *dataset*. En ellas, el valor esperado coincide con el observado. Mientras más cerca esté un punto predicho por alguno de los modelos de la identidad, más exacta es su predicción. Las dos líneas azules marcan la banda de confianza que representa la variabilidad en las predicciones del modelo. Este intervalo muestra el rango donde es probable que se encuentre la verdadera relación entre las variables para un nivel de confianza específico, como el 95%.

Es posible observar cómo, para valores bajos de bailabilidad, los modelos predicen valores más altos para la variable, mientras que, para valores altos, el valor obtenido es menor al esperado.

Para determinar si la hipótesis es válida, se examinaron los coeficientes de los modelos y se evaluaron si las variables *Valence*, *Mode* y *Speechiness* son las mejores predictoras de la bailabilidad de una canción.



Comparación de los coeficientes asociados a cada modelo.

Si bien *Valence* sí es la que presenta una mayor participación en la predicción de la bailabilidad, contrario a lo hipotetizado, no son *Mode* y *Speechiness* quienes se encuentran, junto a *Valence*, en el conjunto de las tres variables mejor predictoras, sino que lo son *Energy* y *Acousticness*, acorde al modelo convencional y de Ridge.

Conclusiones.

A pesar de la aplicación de diversas técnicas, no fue posible obtener conclusiones relevantes de los datos, lo que sugiere que la relación entre las variables musicales es más compleja de lo que inicialmente se supuso y que, posiblemente, requiera de aproximaciones más sofisticadas para un análisis efectivo.

Durante el transcurso del trabajo, se aprendieron conceptos básicos del dominio musical y se aplicaron los conceptos aprendidos en la materia.

Referencias.

- https://escuelademusicalasala.com/ritmo-pulso-compas
- https://espanol.libretexts.org/Humanidades/Musica/Libro:_Apreciaci%C3%B3n_musical_II_(Lumen)/02:_Ritmo_y_Medidor/2.02:_Firma_de_tiempo
- https://espanol.libretexts.org/Humanidades/Musica/Libro:_Entendiendo_la_teor%C3%ADa_b%C3%A1sica_de_la_m%C3%BAsica_(Schmidt-Jones)/02:_Notaci%C3%B3n_-_Tiempo/2.03:_Firma_de_tiempo
- https://es.wikipedia.org/wiki/Sistema_de_notaci%C3%B3n_musical_anglosaj%C3%B3n
- https://creatumusica.art/2022/06/25/notas-musicales/#:~:text=Las%20notas%20musicales%20se%20nombran,alteraciones%20como%20sostenidos%20y%20bemoles