Análisis Estadístico de la predicción de géneros musicales

Dana Acosta, Diana Caro, Santiago Hoyos

Universidad del Rosario
Escuela de Ingeniería, Ciencia y Tecnología

Mayo 26, 2022

Resumen Ejecutivo

A lo largo de la historia se ha ido evidenciando cada vez la importancia de la música en los diferentes aspectos sociales y culturales. De este modo, con el paso de los años surgieron y siguen apareciendo diversos géneros musicales. Así pues, es de particular interés estudiar las implicaciones sobre los géneros musicales haciendo uso de modelos estadísticos.

En este orden de ideas, se quiere realizar una correcta clasificación de los géneros de modo que los usuarios puedan ser capaces de encontrar y descubrir canciones de su completo agrado. De esta manera, también se optimizan los algoritmos en aplicaciones de música vía streaming.

Por último, se estudia la popularidad de los géneros musicales y, asimismo, se predice esta popularidad.

Introducción y descripción del problema

La determinación de géneros musicales puede ser clave para la implementación de nuevos algoritmos y estudios en el campo de la música. Julio Arce, director del Departamento de Musicología de la Universidad Complutense de Madrid, explica a Verne que definir género musical es “algo muy controvertido en la musicología”, porque depende de diferentes factores que, además, cambian en el tiempo. Por ello, en este proyecto se realizó un análisis sobre los factores que se pueden asociar con la música. Esto pueden ser variables como la duración de la canción, nombre, tempo, claves musicales o bailabilidad pueden contribuir a la predicción de los distintos géneros musicales.

Este tipo de estudio puede ser clave para las aplicaciones de reproducción de música vía streaming. Un ejemplo de esto, podría ser la aplicación Spotify, donde uno de sus métodos de clasificación la realiza una inteligencia artificial que toma en cuenta atributos de las canciones como su bailabilidad, felicidad, relajación, entre otros. Con base en lo anterior, realizar una correcta clasificación de los géneros es de suma importancia puesto que que los usuarios deberían ser capaces de encontrar y descubrir canciones de su completo agrado.

Para este análisis estadístico se tendrá en cuenta una base de datos:

- Prediction of music genre.

Esta base de datos se obtuvo de Kaggle, teniendo en cuenta que debía ser viable para el análisis multivariado.

Es importante aclarar que veremos reflejado un planteamiento general y un desarrollo del problema bajo las pautas dadas por y para el curso de Análisis Estadístico de Datos.

La herramienta que se usará para este análisis de datos será R. Un lenguaje diseñado para el estudio analítico de datos, con el que se harán histogramas, gráficas poblacionales y tablas que ayudarán a la descripción en general del problema en cuestión.

Objetivos

1. Analizar las variables para conocer las que indican una verdadera correlación que ayude a la predicción del género de la canción.
2. Hacer uso de los diferentes temas vistos en clase como análisis factorial, Regresión Lineal, PCA y Clustering.
3. Predecir la variable que indica la popularidad de las canciones, usando regresion lineal, con el fin de encontrar el género más popular.
4. Predecir el género musical a partir de valores específicos que nos permiten caracterizarlo, es decir, a partir de unas características predecir de forma aproximada el comportamiento de la variable dataset.

Datos a usar

La base de datos cuenta con los siguientes atributos:

1. **Instance_id:** Identificación única para cada música.
2. **artist_name:** Nombre del artista.
3. **track_name:** Nombre de la canción.
4. **popularity:** Que tan popular es la música.
5. **acousticness:** Acústica.
6. **danceability:** Bailabilidad.
7. **duration_ms:** La duracion de la musica en ms.
8. **energy:** Energía.
9. **instrumentalness:** Instrumentalidad.
10. **key:** Clave de la música.
11. **liveness:**
12. **loudness:**

13. **mode:**
14. **speechiness:**
15. **tempo:**
16. **obtaine_date:**
17. **valence:**
18. **music_genre:**

	instance_id <dbl>	artist_name <chr>	track_name <chr>	popularity <dbl>	acousticness <dbl>	danceability <dbl>	duration_ms <dbl>
1	32894	Röyksopp	Röyksopp's Night Out	27	0.00468	0.652	-1
2	46652	Thievery Corporation	The Shining Path	31	0.01270	0.622	218293
3	30097	Dillon Francis	Hurricane	28	0.00306	0.620	215613
4	62177	Dubloadz	Nitro	34	0.02540	0.774	166875
5	24907	What So Not	Divide & Conquer	32	0.00465	0.638	222369
6	89064	Axel Boman	Hello	47	0.00523	0.755	519468

energy <dbl>	instrumentalness <dbl>	key <chr>	liveness <dbl>	loudness <dbl>	mode <chr>	speechiness <dbl>	tempo <chr>	obtained_date <chr>	valence <dbl>	music_genre <chr>
0.941	0.79200	A#	0.115	-5.201	Minor	0.0748	100.889	4-Apr	0.759	Electronic
0.890	0.95000	D	0.124	-7.043	Minor	0.0300	115.00200000000001	4-Apr	0.531	Electronic
0.755	0.01180	G#	0.534	-4.617	Major	0.0345	127.994	4-Apr	0.333	Electronic
0.700	0.00253	C#	0.157	-4.498	Major	0.2390	128.014	4-Apr	0.270	Electronic
0.587	0.90900	F#	0.157	-6.266	Major	0.0413	145.036	4-Apr	0.323	Electronic
0.731	0.85400	D	0.216	-10.517	Minor	0.0412	?	4-Apr	0.614	Electronic

Por otro lado, se eliminaron las variables no numéricas, por ejemplo, instance_id o key.

1 Parte práctica

Librerías utilizadas

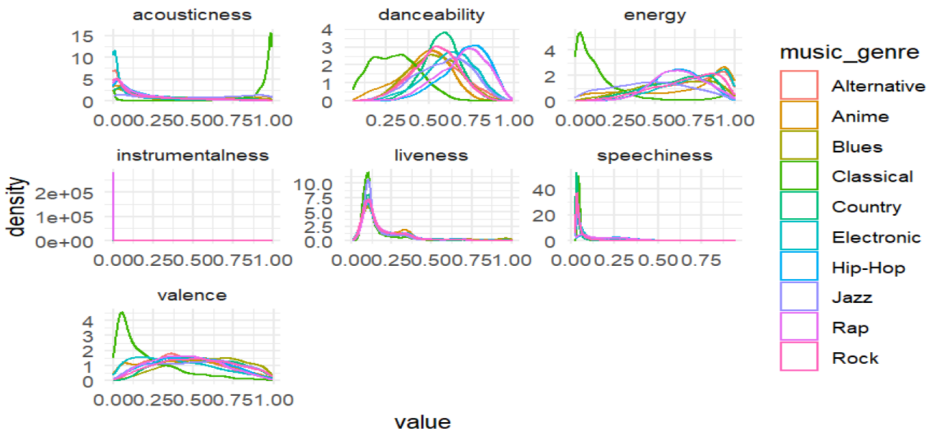
- library(psych)
- library(MASS)
- library(tidyverse)
- library(dplyr)
- library(stats)
- library(FactoMineR)
- library(factoextra)
- library(missMDA)
- library(caret)
- library(magrittr)
- library(plotly)

- library(nortest)
- library(ggridges)
- library(corrplot)
- library(randomForest)
- library(rpart)
- library(rpart.plot)

Análisis de Datos

Se realizaron diversas pruebas de análisis que se ilustrarán a lo largo de este informe.

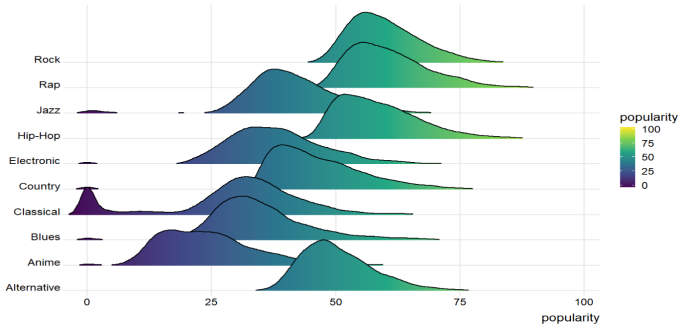
En primer lugar se quiso evaluar si hay variables que caractericen a un género musical (music_genre) en particular:



Al ver las imagenes podriamos decir que la variable energy y valence nos deja ver que la música clásica es la caracterizada por una energia y valence muy alta al inicio.

También, observamos que la variable instrumentalness no aporta nada al estudio del género musical.

Popularidad respecto al género



Así, se evidencia que los géneros más escuchados en los últimos meses presentan niveles de popularidad mayores que temas clásicos, incluso si estos clásicos han sido más populares en el tiempo.

PCA

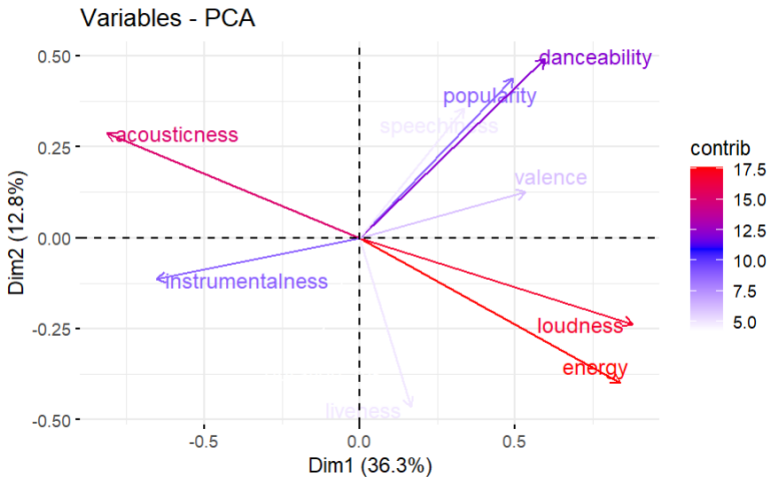
El análisis de componentes principales es una técnica utilizada para describir un conjunto de datos en términos de nuevas variables (≪componentes≫) no correlacionadas. Los componentes se ordenan por la cantidad de varianza original que describen, por lo que la técnica es útil para reducir la dimensionalidad de un conjunto de datos.

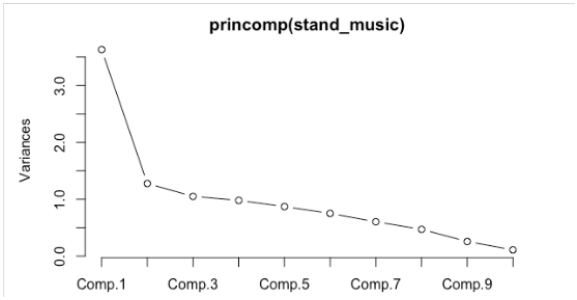
Técnicamente, el PCA busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados. Esta convierte un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables sin correlación lineal llamadas componentes principales.

```
[1] 50000    10
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     1.9054795 1.1297132 1.0248982 0.98955687
Proportion of Variance 0.3630925 0.1276277 0.1050437 0.09792424
Cumulative Proportion  0.3630925 0.4907202 0.5957640 0.69368822
                           Comp.5     Comp.6     Comp.7
Standard deviation     0.93287714 0.86633064 0.77724787
Proportion of Variance 0.08702772 0.07505438 0.06041263
Cumulative Proportion  0.78071593 0.85577031 0.91618295
                           Comp.8     Comp.9    Comp.10
Standard deviation     0.68588302 0.50683936 0.33291458
Proportion of Variance 0.04704449 0.02568913 0.01108343
Cumulative Proportion  0.96322744 0.98891657 1.00000000
```

Con estos resultados tenemos varias cosas por notar, la primera es en cuanto a ”cumulative proportion” vemos que con 6 componentes ya podemos explicar un 85.5% de los datos, y luego, hay acumulación del 96% hasta la componente 8. Sin embargo, la acumulación hasta la componente 6 del 85% no es una mala explicación de nuestros datos.

Además, realizamos la anterior gráfica con el fin de observar cuales son algunas de las variables más importantes. En este caso las líneas rojas son las variables que explican de mejor manera los datos.

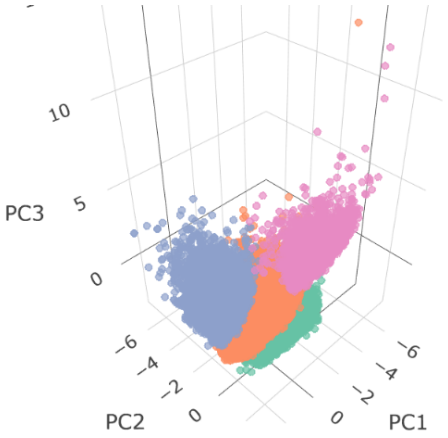




Ahora, observamos la ”grafica del codo” y notamos que el ”codo” se da en la segunda componente.

Clustering

Es la tarea de agrupar objetos por similitud, en grupos o conjuntos de manera que los miembros del mismo grupo tengan características similares. Es la tarea principal de la minería de datos exploratoria y es una técnica común en el análisis de datos estadísticos.



Para realizar la gráfica respectiva a las 4 separaciones que realiza el clustering tuvimos en cuenta los componentes que encontramos en el PCA, donde cada PC1, PC2 Y PC3, corresponde a los primeros 3 componentes respectivamente. Ahora, es claro que no se ve una buena separación de los datos ya que vemos demasiados alejados de los grupos principales y sin embargo clasificados en los mismos.

Regresión

Este método es aplicable en muchas situaciones en las que se estudia la relación entre dos o más variables o predecir un comportamiento, algunas incluso sin relación con la tecnología. En caso de que no se pueda aplicar un modelo de regresión a un estudio, se dice que no hay correlación entre las variables estudiadas.

La regresión decidimos hacerla para predecir popularidad entre los distintos géneros musicales.

```
Residuals:
    Min      1Q  Median      3Q     Max
-58.646  -5.956  -0.739   5.441  57.483

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)              60.92043    0.34423  176.975  < 2e-16 ***
music_genre_Alternative  -8.13600    0.19183  -42.413  < 2e-16 ***
music_genre_Rock          1.39511    0.19198    7.267 3.73e-13 ***
music_genre_Anime       -33.80024    0.19389 -174.331  < 2e-16 ***
music_genre_Blues       -23.14044    0.19387 -119.361  < 2e-16 ***
music_genre_Classical   -26.57177    0.24735 -107.427  < 2e-16 ***
music_genre_Country     -12.24473    0.19185  -63.825  < 2e-16 ***
music_genre_Electronic  -20.13806    0.19341 -104.122  < 2e-16 ***
music_genre_Jazz        -16.69802    0.19888  -83.958  < 2e-16 ***
music_genre_Rap           2.06745    0.19090   10.830  < 2e-16 ***
loudness                  0.18767    0.01498   12.532  < 2e-16 ***
acousticness             -0.78447    0.21891   -3.583 0.000339 ***
energy                   -1.69874    0.35024   -4.850 1.24e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.545 on 49987 degrees of freedom
Multiple R-squared:  0.623,     Adjusted R-squared:  0.6229
F-statistic:  6882 on 12 and 49987 DF,  p-value: < 2.2e-16
```

En los resultados vemos que, todo es estadísticamente significativo a un nivel de confianza del

95%. A un nivel de confianza del 95% el género musical Rock en promedio obtiene 1.39 puntos porcentuales más de popularidad que el resto de los generos.

En este caso, por el Multiple R-squared vemos que el ajuste no es tan bueno, puesto que es del 0.621. En cuanto al p-valor, vemos que este se acerca a 0 (especificamente es 0.00000000000000022) por lo que no podemos aceptar la hipótesis nula de que no hay dependencia entre las variables.

Finalmente, y de las partes más importantes de esta regresián están los coefficients, estos, con la ayuda del intercept nos dejarían (luego) saber la predicción que queremos por cada género respectivo.

```
                           2.5 %     97.5 %
(Intercept)            49.958951  50.489449
music_genreAnime      -26.327719 -25.577481
music_genreBlues      -15.795319 -15.045081
music_genreClassical  -21.283519 -20.533281
music_genreCountry     -4.589319  -3.839081
music_genreElectronic -12.487519 -11.737281
music_genreHip-Hop      7.800281   8.550519
music_genreJazz        -9.670719  -8.920481
music_genreRap          9.898081  10.648319
music_genreRock         9.041881   9.792119
      fit      lwr      upr
1 38.1118 37.84655 38.37705
2 38.1118 37.84655 38.37705
3 38.1118 37.84655 38.37705
4 38.1118 37.84655 38.37705
5 38.1118 37.84655 38.37705
6 38.1118 37.84655 38.37705
      fit    lwr     upr
1 38.1118 19.354 56.8696
2 38.1118 19.354 56.8696
3 38.1118 19.354 56.8696
4 38.1118 19.354 56.8696
5 38.1118 19.354 56.8696
6 38.1118 19.354 56.8696
```

Ahora bien, observamos los resultados de la función de r predict(). Donde vemos que en los intervalos de confianza vs intervalos de predicción hay diferencias bastante importantes.

TEST NORMALIDAD: KOLMOGOROV: LILLIE:

En estadística , la prueba de Lilliefors es una prueba de normalidad basada en la prueba de Kolmogorov-Smirnov. Se utiliza para probar la hipótesis nula de que los datos provienen de una población con distribución normal , cuando la hipótesis nula no especifica qué distribución normal; es decir, no especifica el valor esperado y la varianza de la distribución.

Se quiso hacer uso de este método para el análisis del proyecto. No obstante, para no definir qué probabilidad se tenía que usar, se aplicó el Lillie.test, con el que se consiguió ver que todas las variables tenían un p_valor $<$ 2.2e-16.

```
data:  stand_music$popularity
D = 0.048305, p-value < 2.2e-16

data:  stand_music$acousticness
D = 0.1847, p-value < 2.2e-16

data:  stand_music$danceability
D = 0.024392, p-value < 2.2e-16

data:  stand_music$duration_ms
D = 0.11673, p-value < 2.2e-16

data:  stand_music$energy
D = 0.067108, p-value < 2.2e-16

data:  stand_music$instrumentalness
D = 0.36994, p-value < 2.2e-16

data:  stand_music$liveness
D = 0.2028, p-value < 2.2e-16

data:  stand_music$loudness
D = 0.15332, p-value < 2.2e-16

data:  stand_music$speechiness
D = 0.25484, p-value < 2.2e-16

data:  stand_music$valence
D = 0.041871, p-value < 2.2e-16
```

Como se puede ver, en cada una de las variables numéricas a las que se le aplicó el test, se rechazó la hipótesis nula. Y dado que los datos provienen de la distribución normal, entonces no es posible implementar qda. Por ende, se realizará la prueba de varianzas iguales.

Matriz de Covarianzas

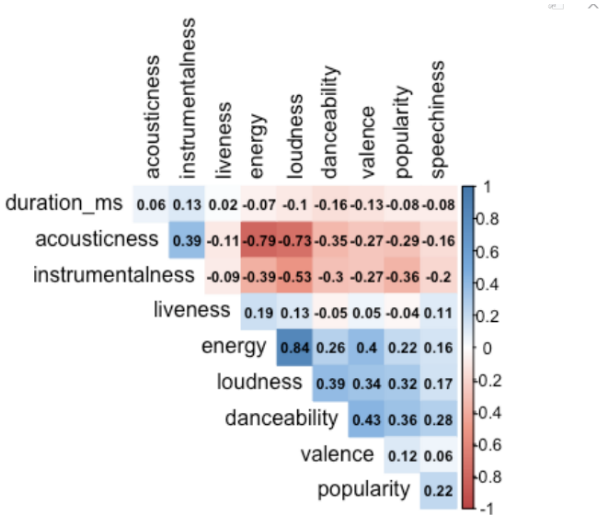
Se realiza la matriz de covarianzas para verificar la factibilidad de usar el método del discriminante lineal, ya que sabemos que necesitamos que las varianzas sean igual para usarlo.

```
                  popularity acousticness danceability
popularity         1.00000000  -0.29045274   0.35641985
acousticness      -0.29045274   1.00000000  -0.34768089
danceability       0.35641985  -0.34768089   1.00000000
duration_ms       -0.08380906   0.06186248  -0.15550694
energy             0.21634534  -0.79124959   0.26393353
instrumentalness  -0.36495980   0.38796964  -0.30127715
liveness          -0.03946831  -0.10921959  -0.05146453
loudness           0.31794092  -0.73040149   0.39085867
speechiness        0.22430947  -0.16337669   0.27976709
valence            0.12491301  -0.27023765   0.43455720
                  duration_ms      energy instrumentalness
popularity        -0.08380906  0.21634534       -0.36495980
acousticness       0.06186248 -0.79124959        0.38796964
danceability      -0.15550694  0.26393353       -0.30127715
duration_ms        1.00000000 -0.06620275        0.12973833
energy            -0.06620275  1.00000000       -0.38972952
instrumentalness   0.12973833 -0.38972952        1.00000000
liveness           0.01991093  0.18673125       -0.09353658
loudness          -0.10250892  0.83839918       -0.52903024
speechiness       -0.08363642  0.15817975       -0.19913675
valence           -0.13057396  0.39631741       -0.27277997
                     liveness   loudness speechiness     valence
popularity       -0.03946831  0.3179409  0.22430947  0.12491301
acousticness     -0.10921959 -0.7304015 -0.16337669 -0.27023765
danceability     -0.05146453  0.3908587  0.27976709  0.43455720
duration_ms       0.01991093 -0.1025089 -0.08363642 -0.13057396
energy            0.18673125  0.8383992  0.15817975  0.39631741
instrumentalness -0.09353658 -0.5290302 -0.19913675 -0.27277997
liveness          1.00000000  0.1265044  0.11447542  0.05210979
loudness          0.12650437  1.0000000  0.16557835  0.34066380
speechiness       0.11447542  0.1655783  1.00000000  0.05984699
valence           0.05210979  0.3406638  0.05984699  1.00000000
```

Vemos que la diagonal principal es de 1's, es decir las varianzas son 1, por lo que confirmamos que podemos usar el método.

Matriz de Correlación

Realizamos la matriz de correlaciones con el fin de aportar a nuestro análisis al revisar los coeficientes de conexión entre los factores. Cada componente de la matriz nos muestra la conexión entre los dos factores.



Energy y loudness son las que tienen una correlación positiva. Energy es una medida de la intensidad. Las características que contribuyen al cálculo de este valor incluyen rango dinámico, volumen percibido, timbre, frecuencia de inicio y entropía general. Por lo que ciertas características que afectan al cálculo de esta variable, también formarán parte del cálculo de loudness, que es el volumen general de un canción.

acousticness y energy: Tienen una alta correlación negativa. Si una canción tiene un nivel de acousticness alto, tendrá menos probabilidad de contener sonidos no acústicos y por lo tanto el timbre o el rango dinámico estará más limitado. Esto afectará a los niveles de energy, que serán inferiores.

Anova

Es el test estadístico a emplear cuando se desea comparar las medias de dos o más grupos. Esta técnica puede generalizarse también para estudiar los posibles efectos de los factores sobre la varianza de una variable.

Se realizó un anova con popularity como variable independiente y music_genre como variable respuesta con el fin de saber cuál es el género más popular.

```
               Df  Sum Sq Mean Sq F value Pr(>F)
music_genre     9 7499805  833312    9100 <2e-16 ***
Residuals   49990 4577655      92
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observamos en los resultados que, existen diferencias estadísticamente significativas entre los géneros musicales ($F(2, 49990) = 9100$ y p_valor es menor que <0.05). Debido a que, ANOVA solo nos muestra que existe una diferencia significativa entre las medias de al menos dos grupos y rechazamos la hipótesis nuña, usaremos el test TukeyHDS que nos mostrará cual es el género que difiere en sus medias.

```
Fit: aov(formula = popularity ~ music_genre, data = musicaa)

$music_genre
                              diff         lwr         upr     p adj
Anime-Alternative       -25.9526 -26.5580845 -25.3471155 0.0000000
Blues-Alternative       -15.4202 -16.0256845 -14.8147155 0.0000000
Classical-Alternative   -20.9084 -21.5138845 -20.3029155 0.0000000
Country-Alternative      -4.2142  -4.8196845  -3.6087155 0.0000000
Electronic-Alternative  -12.1124 -12.7178845 -11.5069155 0.0000000
Hip-Hop-Alternative       8.1754   7.5699155   8.7808845 0.0000000
Jazz-Alternative         -9.2956  -9.9010845  -8.6901155 0.0000000
Rap-Alternative          10.2732   9.6677155  10.8786845 0.0000000
Rock-Alternative          9.4170   8.8115155  10.0224845 0.0000000
Blues-Anime              10.5324   9.9269155  11.1378845 0.0000000
Classical-Anime           5.0442   4.4387155   5.6496845 0.0000000
Country-Anime            21.7384  21.1329155  22.3438845 0.0000000
Electronic-Anime         13.8402  13.2347155  14.4456845 0.0000000
Hip-Hop-Anime            34.1280  33.5225155  34.7334845 0.0000000
Jazz-Anime               16.6570  16.0515155  17.2624845 0.0000000
Rap-Anime                36.2258  35.6203155  36.8312845 0.0000000
Rock-Anime               35.3696  34.7641155  35.9750845 0.0000000
Classical-Blues          -5.4882  -6.0936845  -4.8827155 0.0000000
Country-Blues            11.2060  10.6005155  11.8114845 0.0000000
Electronic-Blues          3.3078   2.7023155   3.9132845 0.0000000
Hip-Hop-Blues            23.5956  22.9901155  24.2010845 0.0000000
Jazz-Blues                6.1246   5.5191155   6.7300845 0.0000000
Rap-Blues                25.6934  25.0879155  26.2988845 0.0000000
Rock-Blues               24.8372  24.2317155  25.4426845 0.0000000
Country-Classical        16.6942  16.0887155  17.2996845 0.0000000
Electronic-Classical      8.7960   8.1905155   9.4014845 0.0000000
Hip-Hop-Classical        29.0838  28.4783155  29.6892845 0.0000000
Jazz-Classical           11.6128  11.0073155  12.2182845 0.0000000
Rap-Classical            31.1816  30.5761155  31.7870845 0.0000000
Rock-Classical           30.3254  29.7199155  30.9308845 0.0000000
Electronic-Country       -7.8982  -8.5036845  -7.2927155 0.0000000
Hip-Hop-Country          12.3896  11.7841155  12.9950845 0.0000000
Jazz-Country             -5.0814  -5.6868845  -4.4759155 0.0000000
Rap-Country              14.4874  13.8819155  15.0928845 0.0000000
Rock-Country             13.6312  13.0257155  14.2366845 0.0000000
Hip-Hop-Electronic       20.2878  19.6823155  20.8932845 0.0000000
Jazz-Electronic           2.8168   2.2113155   3.4222845 0.0000000
Rap-Electronic           22.3856  21.7801155  22.9910845 0.0000000
Rock-Electronic          21.5294  20.9239155  22.1348845 0.0000000
Jazz-Hip-Hop            -17.4710 -18.0764845 -16.8655155 0.0000000
Rap-Hip-Hop               2.0978   1.4923155   2.7032845 0.0000000
Rock-Hip-Hop              1.2416   0.6361155   1.8470845 0.0000000
Rap-Jazz                 19.5688  18.9633155  20.1742845 0.0000000
Rock-Jazz                18.7126  18.1071155  19.3180845 0.0000000
Rock-Rap                 -0.8562  -1.4616845  -0.2507155 0.0003282
```

Vemos que los géneros más alejados en popularidad son Hip-Hop-Anime con Hip-Hop promediando un 34.12 más de popularidad que Anime, luego, Rap-Anime con Rap 36.22 más popular que Anime, y Rock-Anime con Rock 35.36 más popular que Anime. Además, vemos que el valor más pequeño en diferencia es entre Anime-Alternative, lo que nos deja saber que su popularidad es muy parecida y es muy baja.

Con estos resultados podemos ver que los géneros más populares en nuestro estudio son Rock y Rap ya que además vemos que difieren un -0.85 que no es mucho y que Rock sobrepasa la popularidad de Rap por muy poco.

LDA

El Análisis Discriminante Lineal o Linear Discrimiant Analysis (LDA) es un método de clasificación supervisado de variables cualitativas en el que dos o más grupos son conocidos a priori y nuevas observaciones se clasifican en uno de ellos en función de sus características.

Es una alternativa a la regresión logística cuando la variable cualitativa tiene más de dos niveles. Si bien existen extensiones de la regresión logística para múltiples clases, el LDA presenta una serie de ventajas:

- Si las clases están bien separadas, los parámetros estimados en el modelo de regresión logística son inestables. El método de LDA no sufre este problema.
- Si el número de observaciones es bajo y la distribución de los predictores es aproximadamente normal en cada una de las clases, LDA es más estable que la regresión logística.

```
Prior probabilities of groups:
Alternative       Anime       Blues   Classical     Country  Electronic
 0.09982857  0.09965714  0.10180000  0.09942857  0.10051429  0.10014286
    Hip-Hop        Jazz         Rap        Rock
 0.10042857  0.09917143  0.09900000  0.10002857

Group means:
            acousticness danceability duration_ms instrumentalness speechiness
Alternative    0.1659237    0.5401631    211732.8      0.060054287  0.08933111
Anime          0.2870002    0.4706248    207830.9      0.280732420  0.06448572
Blues          0.3212445    0.5270250    227854.9      0.092679829  0.06084985
Classical      0.8757021    0.3048555    278219.1      0.605848518  0.05116810
Country        0.2663943    0.5750748    194651.6      0.005572174  0.04947536
Electronic     0.1223120    0.6194708    244576.7      0.349853874  0.09994765
Hip-Hop        0.1793233    0.7167477    199032.3      0.010754898  0.20903684
Jazz           0.4952145    0.5845745    238132.1      0.352240510  0.07339663
Rap            0.1688046    0.6970762    196370.7      0.008632884  0.18599945
Rock           0.1889347    0.5408865    212019.5      0.054393534  0.05318069
            popularity
Alternative   50.28363
Anime         24.16858
Blues         34.83497
Classical     29.01609
Country       46.00114
Electronic    38.12126
Hip-Hop       58.44011
Jazz          40.96111
Rap           60.48975
Rock          59.76921

Coefficients of linear discriminants:
                        LD1           LD2           LD3           LD4
acousticness      1.031697e+00  2.706092e+00 -1.439592e-01 -1.949898e+00
danceability     -2.485943e+00 -1.823616e+00 -3.468141e+00 -1.416053e-01
duration_ms       4.863177e-07  8.896205e-07 -1.048170e-06  1.033339e-06
instrumentalness  9.556260e-01  9.029999e-01 -1.951703e+00  2.973558e+00
speechiness      -2.986127e+00  2.071683e+00 -8.245856e+00 -3.769910e+00
popularity       -7.838570e-02  5.897602e-02  2.465211e-02  2.918248e-02
                        LD5           LD6
acousticness     -1.577873e+00  9.674892e-02
danceability     -5.532809e+00  1.333181e-02
duration_ms      -8.625411e-07 -7.625433e-06
instrumentalness  1.650750e-01  9.482705e-01
speechiness       6.462011e+00 -4.876449e-01
popularity        3.653960e-03  3.846883e-03

Proportion of trace:
   LD1    LD2    LD3    LD4    LD5    LD6
0.7144 0.1353 0.0992 0.0307 0.0189 0.0015
```

Debido a que las varianzas son iguales se decidió usar el método de discriminante lineal a pesar de la falta de normalidad, además, otra razón para usar este método es que el clustering no dio el resultado que esperado.

Por otro lado, se usaron las variables numéricas más significativas para el género, junto con esta variable.

Vemos que el método calcula las probabilidades donde cada género se acerca o es 0.1. Luego, vemos el promedio de cada género por variables y los coeficientes lineales importantes para la construcción de nuestros discriminantes lineales a partir de los datos train.

Predicción LDA

Realizamos la predicción del modelo:

Predicted	Actual Alternative	Anime	Blues	Classical	Country	Electronic	Hip-Hop	Jazz
Alternative	1213	129	210	59	383	243	217	85
Anime	14	1660	549	200	31	246	2	88
Blues	100	748	1640	170	747	426	5	498
Classical	12	668	121	2871	7	76	0	377
Country	771	80	365	30	1466	269	221	475
Electronic	184	124	220	51	42	1685	27	761
Hip-Hop	343	6	32	1	92	205	1807	181
Jazz	126	56	281	88	148	222	14	882
Rap	95	2	10	0	18	33	801	20
Rock	636	15	135	10	584	100	421	104

Predicted	Actual Rap	Rock
Alternative	228	522
Anime	2	4
Blues	3	8
Classical	0	16
Country	145	306
Electronic	12	51
Hip-Hop	1451	42
Jazz	9	59
Rap	871	154
Rock	744	2339

Con este resultado podemos ver los valores predictores para cada género musical.

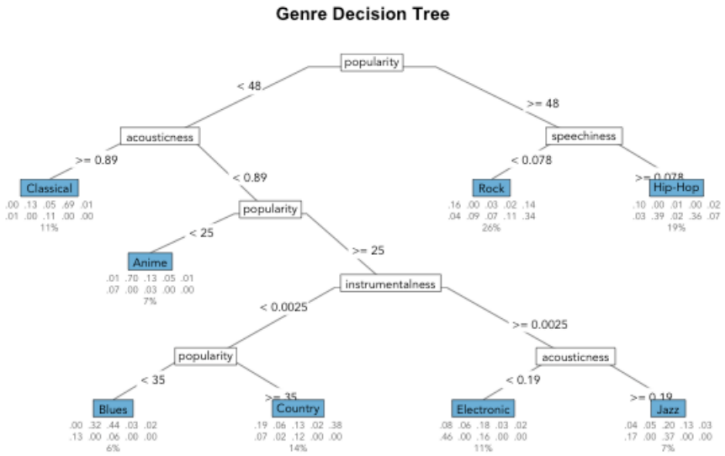
Particionamiento recursivo

El particionamiento recursivo es una técnica estadística de análisis multivariante. Su objetivo es el de construir árboles de decisión que modelen la influencia de una serie de variables explicativas sobre la variable objetivo de un estudio estadístico.

Los modelos construidos con esta técnica rivalizan con otros más tradicionales de la estadística —por ejemplo, regresiones logísticas— o de la inteligencia artificial, como los basados en redes neuronales.

Árbol con partición de datos train, usada también en el LDA

Se usa este método con el fin de ver los resultados de lda mucho más claros y encontrar nuestras variables más importantes a la hora de clasificar un género musical. Esta sección se realiza con el conjunto de datos trai, es decir, con el 70% de los datos.



Con los resultados de la gráfica vemos que popularity es una variable muy importante para clasificar el género, además de acousticness, speechiness y instrumentalness una vez se haya clasificado popularidad y acousticness de una canción.

Matriz de confusión para la partición de datos train:

```
Confusion Matrix and Statistics

            dec_tree_result
             Alternative Anime Blues Classical Country
Alternative            0    14     1        15     936
Anime                  0  1652   636       487     328
Blues                  0   294   871       184     648
Classical              0    95    61      2517     103
Country                0    42    44        48    1908
Electronic             0   167   256        25     343
Hip-Hop                0     1     2         0     118
Jazz                   0    74   107       417     601
Rap                    0     3     3         0      22
Rock                   0     6     8         0      20
            dec_tree_result
             Electronic Hip-Hop Jazz  Rap Rock
Alternative        320     703   82    0 1437
Anime              234      24  138    0   40
Blues              737      51  472    0  294
Classical          131       8  331    0  198
Country             80     104   78    0 1259
Electronic        1753     177  395    0  371
Hip-Hop             11    2588    2    0  764
Jazz               660     166  894    0  567
Rap                  2    2471    1    0  988
Rock                12     492    2    0 2906
```

- Accuracy : 0.4311
- P-Value : $<$ 2.2e-16
- Sensibilidad, especificidad y los valores positivos y negativos que se predicen para cada genero.

Por ejemplo, para los tres primeros generos Alternative, Anime y Blues tenemos los siguientes valores respectivamente:

Sensitivity	NA	0.71921	0.42315
Specificity	0.8992	0.94116	0.92008
Pos Pred Value	NA	0.45444	0.24079
Neg Pred Value	NA	0.98007	0.96381

Table 1: Fruta disponible

El ratio de aciertos general es de 0.4361

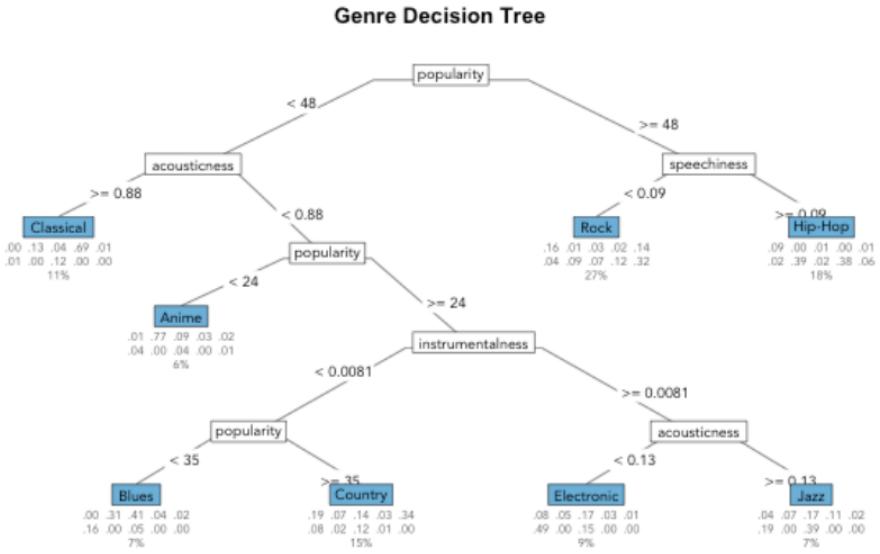
Para el test group, el accuracy sube (aunque no mucho) y el valor p no cambia t en cuanto a los valores de sensibilidad, especificidad y positivos y negativos, no se ve un cambio grande.

La prediccion es del 46%

El radio de aciertos general es de 0.4361.

Árbol para partición de datos test

Para esta sección usamos el 30% de datos restantes, es decir la partición test.



Vemos que los resultados son similares a los del árbol anterior, exceptuando los pesos de las aristas, es decir, los pesos o características que debe presentar cada canción para ser clasificada; como por ejemplo si la popularidad es mayor que 48 y su speechiness es menor que 0.09 se trata de una canción de Rock.

Matriz de confusión para la partición de datos test:

```
Confusion Matrix and Statistics

              dec_tree_result2I
               Alternative Anime Blues Classical Country Electronic Hip-Hop Jazz
  Alternative            0     7     1         8     439        102     242   46
  Anime                  0   649   313       210     157         72      10   76
  Blues                  0    77   423        63     322        230      16  183
  Classical              0    25    37      1125      66         42       6  120
  Country                0    16    17        19     799         18      37   27
  Electronic             0    35   167        14     187        649      65  212
  Hip-Hop                0     1     2         0      54          5    1055    1
  Jazz                   0    31    56       202     274        205      60  431
  Rap                    0     2     2         0      12          1    1030    1
  Rock                   0     5     5         0       8          4     163    1
              dec_tree_result2I
                Rap Rock
  Alternative     0  661
  Anime           0   25
  Blues           0  123
  Classical       0   99
  Country         0  549
  Electronic      0  166
  Hip-Hop         0  367
  Jazz            0  270
  Rap             0  487
  Rock            0 1313
```

Para el test group, el accuracy sube (aunque no mucho) y el valor p no cambia t en cuanto a los valores de sensibilidad, especificidad y positivos y negativos, no se ve un cambio grande. La prediccion es del 46%

Conclusiones

En este punto del proyecto y respecto a los resultados obtenidos podemos concluir que en nuestro caso, los géneros más populares son Rock y Rap. Además, gracias a la regresión lineal realizada podemos predecir que el género más importantes será Rock.

Adicionalmente, respecto a nuestro objetivo principal del proyecto, es decir, clasificar los géneros musicales gracias a las variables que caractericen a music_genre pudimos observar que popularity es una variable muy importante junto con acousticness y speechiness. Además, de instrumentalness una vez se haya clasificado la canción según su popularidad y acousticness. Logramos clasificar los géneros Classical, Anime, Blues, Country, Electronic, Jazz, Rock y Hip-Hop.

References

[1] Colaboradores de Wikipedia. (2021, 15 octubre). Análisis de componentes principales. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/An%C3%A1lisisdecomponentesprincipales

[2] Colaboradores de Wikipedia. (2022, 15 mayo). Análisis de grupos. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/An%C3%A1lisisdegrupos

[3] Colaboradores de Wikipedia. (2022a, marzo 26). Regresión lineal. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/Regresi%C3%B3nlineal

[4] Colaboradores de Wikipedia. (2020, 22 octubre). Prueba de Lilliefors. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/PruebadeLilliefors

[5] Prediction of music genre. (2021, 2 noviembre). Kaggle. https://www.kaggle.com/datasets/vicsuperman/prediction-of-music-genre

[6] Rodrigo, J. A. (s. f.). ANOVA análisis de varianza para comparar múltiples medias. Ciencia de Datos. https://www.cienciadedatos.net/documentos/19anova

[7] Rodrigo, J. A. (s. f.-a). Análisis discriminante lineal (LDA) y análisis discriminante cuadrático (QDA). Ciencia de Datos. https://www.cienciadedatos.net/documentos/28lineardiscriminantanalysisldayquadraticdiscriminantanalysisqda

[8] https://rstudio-pubs-static.s3.amazonaws.com/71622192850e1ae9224bb0b6f0e2a58b42f9b4.html63decisiontree