gera_Dataset_predicaoTop200

June 26, 2022

0.1 Notebook para criar o dataset a ser utilizado na pergunta preditiva número 1

Prever se uma música alcançará o Top 200 Global

[ ]:
```
import pandas as pd
import numpy as np
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[ ]:
```
path = '/content/drive/My Drive/INF1032 - Spotify/Dados/Consolidados/'

df_musicas_ruins = pd.read_csv(path+"musicas_ruins.csv",index_col="Unnamed: 0")
df_musicas_boas = pd.read_csv(path+"musicas_boas.csv",index_col="Unnamed: 0")

#Queremos um dataset final balanceado
print("Quantidade de músicas em cada dataset:\n%d boas e %d␣
 ↪ruins"%(df_musicas_ruins.shape[0],df_musicas_boas.shape[0]))
```

```
Quantidade de músicas em cada dataset:
3926 boas e 5156 ruins
```

[ ]:
```
#Verificando se os dataset tem exatamente as mesmas colunas
var_ruins = df_musicas_ruins.columns
var_boas = df_musicas_boas.columns

var_ruins.sort_values()
var_boas.sort_values()

if np.array_equal(var_ruins,var_boas):
  print("Mesmas colunas")

print(var_ruins.values)
print(var_boas.values)
```

```
Mesmas colunas
['danceability' 'energy' 'key' 'loudness' 'mode' 'speechiness'
 'acousticness' 'instrumentalness' 'liveness' 'valence' 'tempo' 'type'
 'id' 'uri' 'track_href' 'analysis_url' 'duration_ms' 'time_signature'
 'nome' 'data_lancamento' 'Popularidade Musica' 'Artista' 'ano_lancamento'
 'mes_lancamento' 'dia_semana_lancamento' 'Popularidade Artista'
 'Seguidores' 'Estilos']
['danceability' 'energy' 'key' 'loudness' 'mode' 'speechiness'
 'acousticness' 'instrumentalness' 'liveness' 'valence' 'tempo' 'type'
 'id' 'uri' 'track_href' 'analysis_url' 'duration_ms' 'time_signature'
 'nome' 'data_lancamento' 'Popularidade Musica' 'Artista' 'ano_lancamento'
 'mes_lancamento' 'dia_semana_lancamento' 'Popularidade Artista'
 'Seguidores' 'Estilos']
```

[ ]:
```
df_musicas_ruins.head(10)
```

[ ]:

	danceability	energy	key	loudness	mode	speechiness	acousticness	\
0	0.371	0.609	11	-5.524	1	0.0295	0.386000	
1	0.552	0.368	4	-6.804	1	0.0383	0.816000	
2	0.500	0.824	4	-5.846	0	0.0388	0.000160	
3	0.681	0.655	0	-8.642	1	0.0257	0.042500	
4	0.790	0.687	7	-7.571	1	0.1010	0.014900	
5	0.278	0.676	9	-8.821	1	0.0371	0.000423	
6	0.492	0.534	7	-8.361	1	0.0333	0.292000	
7	0.499	0.916	9	-4.785	0	0.0501	0.000391	
8	0.409	0.920	11	-5.439	0	0.1200	0.007190	
9	0.755	0.842	2	-7.293	0	0.1290	0.388000	

	instrumentalness	liveness	valence	...	nome	\
0	0.000000	0.1040	0.426	...	I'll Be	
1	0.000000	0.0930	0.486	...	Used To Love - Acoustic Version	
2	0.001610	0.0916	0.539	...	Come As You Are	
3	0.000000	0.0699	0.901	...	Super Trouper	
4	0.000319	0.3230	0.628	...	What I Got	
5	0.000899	0.1360	0.494	...	Take Me Out	
6	0.000000	0.2200	0.892	...	Big Yellow Taxi	
7	0.001170	0.1540	0.251	...	Save Me	
8	0.001050	0.0823	0.370	...	Cum on Feel the Noize	
9	0.541000	0.1050	0.798	...	Could Heaven Ever Be Like This	

	data_lancamento	Popularidade Musica	Artista	ano_lancamento	\
0	1997-01-01	68	Edwin McCain	1997.0	
1	2020-02-14	48	Martin Garrix	2020.0	
2	1991-09-26	12	Nirvana	1991.0	
3	2018-07-13	61	Cher	2018.0	
4	1996-07-30	0	Sublime	1996.0	
5	2004-01-01	9	Franz Ferdinand	2004.0	
6	1973-11-16	31	Bob Dylan	1973.0	

```
7      2010-07-23                  54  Avenged Sevenfold         2010.0
8      1983-01-01                  72         Quiet Riot         1983.0
9      1977-01-01                  58      Idris Muhammad         1977.0

  mes_lancamento  dia_semana_lancamento  Popularidade Artista Seguidores  \
0            1.0                    2.0                    52     275100
1            2.0                    4.0                    76   15567599
2            9.0                    3.0                    80   15532964
3            7.0                    4.0                    67    2570384
4            7.0                    1.0                    69    2250355
5            1.0                    3.0                    64    2259453
6           11.0                    4.0                    70    5725786
7            7.0                    4.0                    72    5415058
8            1.0                    5.0                    58    1447370
9            1.0                    5.0                    42      58936

                                        Estilos
0                        ['neo mellow', 'pop rock']
1  ['dance pop', 'dutch edm', 'edm', 'pop', 'pop ...
2               ['grunge', 'permanent wave', 'rock']
3  ['dance pop', 'hollywood', 'new wave pop', 'pop']
4      ['reggae fusion', 'ska mexicano', 'ska punk']
5  ['alternative rock', 'dance rock', 'indie rock...
6  ['classic rock', 'country rock', 'folk', 'folk...
7                  ['alternative metal', 'nu metal']
8  ['album rock', 'classic rock', 'glam metal', '...
9  ['funk', 'jazz funk', 'new orleans funk', 'sou...

[10 rows x 28 columns]
```

[ ]:
```
df_musicas_boas.head(10)
```

[ ]:
```
   danceability  energy  key  loudness  mode  speechiness  acousticness  \
0         0.520   0.731    6    -5.338     0       0.0557        0.3420
1         0.905   0.563    8    -6.135     1       0.1020        0.0254
2         0.761   0.525   11    -6.900     1       0.0944        0.4400
3         0.591   0.764    1    -5.484     1       0.0483        0.0383
4         0.756   0.697    8    -6.377     1       0.0401        0.1820
5         0.728   0.783   11    -4.424     0       0.2660        0.2370
6         0.795   0.800    1    -6.320     1       0.0309        0.0354
7         0.741   0.691   10    -7.395     0       0.0672        0.0221
8         0.812   0.736    4    -5.421     0       0.0833        0.1520
9         0.870   0.548   10    -5.253     0       0.0770        0.0924

   instrumentalness  liveness  valence  ...  \
0          0.001010    0.3110    0.662  ...
1          0.000010    0.1130    0.324  ...
2          0.000007    0.0921    0.531  ...
```

```
3          0.000000    0.1030    0.478  ...
4          0.000000    0.3330    0.956  ...
5          0.000000    0.4340    0.555  ...
6          0.000073    0.0915    0.934  ...
7          0.000000    0.0476    0.892  ...
8          0.002540    0.0914    0.396  ...
9          0.000046    0.0534    0.832  ...

                                                 nome data_lancamento  \
0                                           As It Was      2022-03-31
1                                         First Class      2022-04-08
2                                          Heat Waves      2020-08-06
3                              STAY (with Justin Bieber)      2021-07-23
4                             Bam Bam (feat. Ed Sheeran)      2022-04-07
5  Enemy (with JID) - from the series Arcane Leag...      2021-09-03
6                               Cold Heart - PNAU Remix      2021-10-22
7                    INDUSTRY BABY (feat. Jack Harlow)      2021-09-17
8                                            Envolver      2022-04-12
9                               Una Noche en Medellín      2022-01-21

  Popularidade Musica          Artista ano_lancamento mes_lancamento  \
0                 100     Harry Styles         2022.0            3.0
1                  94      Jack Harlow         2022.0            4.0
2                  90    Glass Animals         2020.0            8.0
3                  88    The Kid LAROI         2021.0            7.0
4                  85   Camila Cabello         2022.0            4.0
5                  84  Imagine Dragons         2021.0            9.0
6                  85       Elton John         2021.0           10.0
7                  85        Lil Nas X         2021.0            9.0
8                  79           Anitta         2022.0            4.0
9                  92          Cris Mj         2022.0            1.0

   dia_semana_lancamento  Popularidade Artista Seguidores  \
0                    3.0                    94   21444145
1                    4.0                    86    2247792
2                    3.0                    80    2960684
3                    4.0                    83    3778109
4                    3.0                    83   27026106
5                    4.0                    86   40637595
6                    4.0                    82    9483990
7                    4.0                    82   11174940
8                    1.0                    80   11762155
9                    4.0                    78     409792

                                          Estilos
0                                         ['pop']
1  ['deep underground hip hop', 'kentucky hip hop...
```

```
2        ['gauze pop', 'indietronica', 'shiver pop']
3                             ['australian hip hop']
4             ['dance pop', 'pop', 'post-teen pop']
5                            ['modern rock', 'rock']
6         ['glam rock', 'mellow gold', 'piano rock']
7                           ['lgbtq+ hip hop', 'pop']
8  ['funk carioca', 'funk pop', 'pagode baiano', ...
9                ['mambo chileno', 'urbano chileno']

[10 rows x 28 columns]
```

0.2 Coluna de classificação

As músicas que vieram do spotify Charts e que, portanto, apareceram no Top 200 Global, terão valor 1 na coluna *Top*.

As músicas que não apareceram nesse top global terão valor 0 nessa coluna.

[ ]:
```
df_musicas_ruins["Top"] = 0
df_musicas_boas["Top"] = 1
```

[ ]:
```
df_musicas_ruins.head(10)
```

[ ]:
```
   danceability  energy  key  loudness  mode  speechiness  acousticness  \
0         0.371   0.609   11    -5.524     1       0.0295      0.386000
1         0.552   0.368    4    -6.804     1       0.0383      0.816000
2         0.500   0.824    4    -5.846     0       0.0388      0.000160
3         0.681   0.655    0    -8.642     1       0.0257      0.042500
4         0.790   0.687    7    -7.571     1       0.1010      0.014900
5         0.278   0.676    9    -8.821     1       0.0371      0.000423
6         0.492   0.534    7    -8.361     1       0.0333      0.292000
7         0.499   0.916    9    -4.785     0       0.0501      0.000391
8         0.409   0.920   11    -5.439     0       0.1200      0.007190
9         0.755   0.842    2    -7.293     0       0.1290      0.388000

   instrumentalness  liveness  valence  ...  data_lancamento  \
0          0.000000    0.1040    0.426  ...      1997-01-01
1          0.000000    0.0930    0.486  ...      2020-02-14
2          0.001610    0.0916    0.539  ...      1991-09-26
3          0.000000    0.0699    0.901  ...      2018-07-13
4          0.000319    0.3230    0.628  ...      1996-07-30
5          0.000899    0.1360    0.494  ...      2004-01-01
6          0.000000    0.2200    0.892  ...      1973-11-16
7          0.001170    0.1540    0.251  ...      2010-07-23
8          0.001050    0.0823    0.370  ...      1983-01-01
9          0.541000    0.1050    0.798  ...      1977-01-01

  Popularidade Musica            Artista ano_lancamento mes_lancamento  \
0                  68       Edwin McCain         1997.0            1.0
```

```
1                  48      Martin Garrix         2020.0            2.0
2                  12            Nirvana         1991.0            9.0
3                  61               Cher         2018.0            7.0
4                   0            Sublime         1996.0            7.0
5                   9    Franz Ferdinand         2004.0            1.0
6                  31          Bob Dylan         1973.0           11.0
7                  54  Avenged Sevenfold         2010.0            7.0
8                  72         Quiet Riot         1983.0            1.0
9                  58     Idris Muhammad         1977.0            1.0

  dia_semana_lancamento  Popularidade Artista  Seguidores  \
0                   2.0                    52      275100
1                   4.0                    76    15567599
2                   3.0                    80    15532964
3                   4.0                    67     2570384
4                   1.0                    69     2250355
5                   3.0                    64     2259453
6                   4.0                    70     5725786
7                   4.0                    72     5415058
8                   5.0                    58     1447370
9                   5.0                    42       58936

                                        Estilos Top
0                       ['neo mellow', 'pop rock']   0
1  ['dance pop', 'dutch edm', 'edm', 'pop', 'pop ...   0
2              ['grunge', 'permanent wave', 'rock']   0
3  ['dance pop', 'hollywood', 'new wave pop', 'pop']   0
4      ['reggae fusion', 'ska mexicano', 'ska punk']   0
5  ['alternative rock', 'dance rock', 'indie rock...   0
6  ['classic rock', 'country rock', 'folk', 'folk...   0
7                  ['alternative metal', 'nu metal']   0
8  ['album rock', 'classic rock', 'glam metal', '...   0
9  ['funk', 'jazz funk', 'new orleans funk', 'sou...   0

[10 rows x 29 columns]
```

[ ]:
```
df_musicas_boas.head(10)
```

[ ]:
```
   danceability  energy  key  loudness  mode  speechiness  acousticness  \
0         0.520   0.731    6    -5.338     0       0.0557        0.3420
1         0.905   0.563    8    -6.135     1       0.1020        0.0254
2         0.761   0.525   11    -6.900     1       0.0944        0.4400
3         0.591   0.764    1    -5.484     1       0.0483        0.0383
4         0.756   0.697    8    -6.377     1       0.0401        0.1820
5         0.728   0.783   11    -4.424     0       0.2660        0.2370
6         0.795   0.800    1    -6.320     1       0.0309        0.0354
7         0.741   0.691   10    -7.395     0       0.0672        0.0221
8         0.812   0.736    4    -5.421     0       0.0833        0.1520
```

```
9         0.870   0.548   10    -5.253     0       0.0770        0.0924

   instrumentalness  liveness  valence  ...  data_lancamento  \
0          0.001010    0.3110    0.662  ...       2022-03-31
1          0.000010    0.1130    0.324  ...       2022-04-08
2          0.000007    0.0921    0.531  ...       2020-08-06
3          0.000000    0.1030    0.478  ...       2021-07-23
4          0.000000    0.3330    0.956  ...       2022-04-07
5          0.000000    0.4340    0.555  ...       2021-09-03
6          0.000073    0.0915    0.934  ...       2021-10-22
7          0.000000    0.0476    0.892  ...       2021-09-17
8          0.002540    0.0914    0.396  ...       2022-04-12
9          0.000046    0.0534    0.832  ...       2022-01-21

  Popularidade Musica          Artista ano_lancamento mes_lancamento  \
0                 100     Harry Styles         2022.0            3.0
1                  94      Jack Harlow         2022.0            4.0
2                  90    Glass Animals         2020.0            8.0
3                  88    The Kid LAROI         2021.0            7.0
4                  85   Camila Cabello         2022.0            4.0
5                  84  Imagine Dragons         2021.0            9.0
6                  85       Elton John         2021.0           10.0
7                  85        Lil Nas X         2021.0            9.0
8                  79           Anitta         2022.0            4.0
9                  92          Cris Mj         2022.0            1.0

  dia_semana_lancamento  Popularidade Artista  Seguidores  \
0                   3.0                    94    21444145
1                   4.0                    86     2247792
2                   3.0                    80     2960684
3                   4.0                    83     3778109
4                   3.0                    83    27026106
5                   4.0                    86    40637595
6                   4.0                    82     9483990
7                   4.0                    82    11174940
8                   1.0                    80    11762155
9                   4.0                    78      409792

                                          Estilos Top
0                                         ['pop']   1
1  ['deep underground hip hop', 'kentucky hip hop...   1
2        ['gauze pop', 'indietronica', 'shiver pop']   1
3                           ['australian hip hop']   1
4              ['dance pop', 'pop', 'post-teen pop']   1
5                          ['modern rock', 'rock']   1
6         ['glam rock', 'mellow gold', 'piano rock']   1
7                        ['lgbtq+ hip hop', 'pop']   1
```

```
8  ['funk carioca', 'funk pop', 'pagode baiano', ...   1
9                ['mambo chileno', 'urbano chileno']   1

[10 rows x 29 columns]
```

0.3 Concatenando os dois datasets

[ ]:
```
df_predicao_top200 = pd.concat([df_musicas_boas,df_musicas_ruins],axis=0)
df_predicao_top200 = df_predicao_top200.reset_index(drop=True)
df_predicao_top200
```

[ ]:
```
      danceability  energy  key  loudness  mode  speechiness  acousticness  \
0            0.520   0.731    6    -5.338     0       0.0557       0.34200
1            0.905   0.563    8    -6.135     1       0.1020       0.02540
2            0.761   0.525   11    -6.900     1       0.0944       0.44000
3            0.591   0.764    1    -5.484     1       0.0483       0.03830
4            0.756   0.697    8    -6.377     1       0.0401       0.18200
...            ...     ...  ...       ...   ...          ...           ...
9077         0.609   0.777    9    -7.712     1       0.0636       0.01480
9078         0.631   0.932    5    -4.142     1       0.0354       0.04360
9079         0.481   0.435    4    -8.795     1       0.0321       0.67800
9080         0.522   0.889    1    -4.137     1       0.0461       0.00328
9081         0.613   0.589    0   -10.388     1       0.0458       0.10700

      instrumentalness  liveness  valence  ...  data_lancamento  \
0             0.001010    0.3110    0.662  ...       2022-03-31
1             0.000010    0.1130    0.324  ...       2022-04-08
2             0.000007    0.0921    0.531  ...       2020-08-06
3             0.000000    0.1030    0.478  ...       2021-07-23
4             0.000000    0.3330    0.956  ...       2022-04-07
...                ...       ...      ...  ...              ...
9077          0.074600    0.1530    0.546  ...       2021-10-15
9078          0.137000    0.0918    0.971  ...       1981-08-24
9079          0.000000    0.0928    0.107  ...       2014-01-01
9080          0.000000    0.3450    0.852  ...       2006-01-29
9081          0.000036    0.1140    0.757  ...       2000-01-01

     Popularidade Musica             Artista ano_lancamento mes_lancamento  \
0                    100        Harry Styles         2022.0            3.0
1                     94         Jack Harlow         2022.0            4.0
2                     90       Glass Animals         2020.0            8.0
3                     88       The Kid LAROI         2021.0            7.0
4                     85      Camila Cabello         2022.0            4.0
...                  ...                 ...            ...            ...
9077                  50          Remi Wolf         2021.0           10.0
9078                   0  The Rolling Stones         1981.0            8.0
9079                  60        Taylor Swift         2014.0            1.0
```

```
9080                  60       Arctic Monkeys         2006.0            1.0
9081                  50  Yusuf / Cat Stevens         2000.0            1.0

     dia_semana_lancamento  Popularidade Artista  Seguidores  \
0                      3.0                    94    21444145
1                      4.0                    86     2247792
2                      3.0                    80     2960684
3                      4.0                    83     3778109
4                      3.0                    83    27026106
...                    ...                   ...         ...
9077                   4.0                    65      283640
9078                   0.0                    77    11805172
9079                   2.0                    92    54364596
9080                   6.0                    82    14770606
9081                   5.0                    67     1532558

                                              Estilos Top
0                                             ['pop']   1
1     ['deep underground hip hop', 'kentucky hip hop...   1
2          ['gauze pop', 'indietronica', 'shiver pop']   1
3                              ['australian hip hop']   1
4                ['dance pop', 'pop', 'post-teen pop']   1
...                                                ...  ..
9077            ['indie pop', 'modern alternative pop']   0
9078       ['british invasion', 'classic rock', 'rock']   0
9079                                          ['pop']   0
9080  ['garage rock', 'permanent wave', 'rock', 'she...   0
9081  ['british folk', 'classic rock', 'folk', 'folk...   0

[9082 rows x 29 columns]
```

0.3.1 Não podemos ter música repetida

Caso alguma música considerada boa tiver aparecido no Top 200 Global, a linha da tabela que veio das musicas ruins será eliminada

[ ]:
```
df_predicao_top200 = df_predicao_top200.
 ↪drop_duplicates(subset=["nome","duration_ms"],keep="first")

#Queremos um dataset final balanceado
print("Quantidade de músicas em cada dataset:\n%d boas e %d␣
 ↪ruins"%((df_predicao_top200["Top"]==1).sum(),(df_predicao_top200["Top"]==0).
 ↪sum()))
df_predicao_top200.to_csv(path+"dataset_previsao_charts.csv")
df_predicao_top200
```

```
Quantidade de músicas em cada dataset:
4660 boas e 3500 ruins
```

[ ]:

	danceability	energy	key	loudness	mode	speechiness	acousticness
0	0.520	0.731	6	-5.338	0	0.0557	0.34200
1	0.905	0.563	8	-6.135	1	0.1020	0.02540
2	0.761	0.525	11	-6.900	1	0.0944	0.44000
3	0.591	0.764	1	-5.484	1	0.0483	0.03830
4	0.756	0.697	8	-6.377	1	0.0401	0.18200
...	...	...	...	...	...	...	...
9077	0.609	0.777	9	-7.712	1	0.0636	0.01480
9078	0.631	0.932	5	-4.142	1	0.0354	0.04360
9079	0.481	0.435	4	-8.795	1	0.0321	0.67800
9080	0.522	0.889	1	-4.137	1	0.0461	0.00328
9081	0.613	0.589	0	-10.388	1	0.0458	0.10700

	instrumentalness	liveness	valence	...	data_lancamento
0	0.001010	0.3110	0.662	...	2022-03-31
1	0.000010	0.1130	0.324	...	2022-04-08
2	0.000007	0.0921	0.531	...	2020-08-06
3	0.000000	0.1030	0.478	...	2021-07-23
4	0.000000	0.3330	0.956	...	2022-04-07
...	...	...	...	...	...
9077	0.074600	0.1530	0.546	...	2021-10-15
9078	0.137000	0.0918	0.971	...	1981-08-24
9079	0.000000	0.0928	0.107	...	2014-01-01
9080	0.000000	0.3450	0.852	...	2006-01-29
9081	0.000036	0.1140	0.757	...	2000-01-01

	Popularidade Musica	Artista	ano_lancamento	mes_lancamento
0	100	Harry Styles	2022.0	3.0
1	94	Jack Harlow	2022.0	4.0
2	90	Glass Animals	2020.0	8.0
3	88	The Kid LAROI	2021.0	7.0
4	85	Camila Cabello	2022.0	4.0
...	...	...	...	...
9077	50	Remi Wolf	2021.0	10.0
9078	0	The Rolling Stones	1981.0	8.0
9079	60	Taylor Swift	2014.0	1.0
9080	60	Arctic Monkeys	2006.0	1.0
9081	50	Yusuf / Cat Stevens	2000.0	1.0

	dia_semana_lancamento	Popularidade Artista	Seguidores
0	3.0	94	21444145
1	4.0	86	2247792
2	3.0	80	2960684
3	4.0	83	3778109
4	3.0	83	27026106
...	...	...	...
9077	4.0	65	283640

```
9078                   0.0                    77    11805172
9079                   2.0                    92    54364596
9080                   6.0                    82    14770606
9081                   5.0                    67     1532558

                                                 Estilos Top
0                                                 ['pop']   1
1     ['deep underground hip hop', 'kentucky hip hop...   1
2           ['gauze pop', 'indietronica', 'shiver pop']   1
3                                ['australian hip hop']   1
4                 ['dance pop', 'pop', 'post-teen pop']   1
...                                                 ...  ..
9077            ['indie pop', 'modern alternative pop']   0
9078       ['british invasion', 'classic rock', 'rock']   0
9079                                              ['pop']   0
9080  ['garage rock', 'permanent wave', 'rock', 'she...   0
9081  ['british folk', 'classic rock', 'folk', 'folk...   0

[8160 rows x 29 columns]
```