Mineração de conjuntos frequentes em playlists do Spotify

Raul Araju Korogi Oliveira, Gabriel Lima Nunes

[1] Universidade Federal de Minas Gerais (UFMG)

Resumo. *Este artigo consiste em uma análise de um conjunto de playlists do spotify para reconhecimento de padrões frequentes, utilizando a ferramenta do ChatGPT como auxiliar, com o objetivo de julgar sua capacidade de análise e mineração de dados.*

1. Entendimento do Negócio

Nesse trabalho, buscamos realizar a análise de um banco de dados formado por playlists do spotify, com as músicas presentes em cada uma delas e seus artistas e gêneros. Sob essas playlists, desejamos identificar padrões frequentes de músicas e de artistas, utilizando a ferramenta ChatGPT como suporte, visando analisar a sua capacidade de análise e mineração de dados, e comparando seu desempenho e escolhas ao planejamento prévio.

O ChatGPT foi escolhido como a ferramenta para esse trabalho devido a sua popularidade, sendo uma das inteligências artificiais baseadas em chat mais famosa e uma das responsáveis pelo uso tão abrangente dessas ferramentas. Outras vantagens dessa ferramenta que influenciaram sua decisão incluem a presença de uma versão gratuita e o tamanho da base de dados com a qual foi treinado.

Inicialmente, informamos para o ChatGPT 7.1 que possuíamos um trabalho de mineração de dados e precisaríamos da ajuda dele para a sua confecção, pedindo para ele informar que dados ele precisaria para poder nos auxiliar. Ele pediu o nosso objetivo com o trabalho, que base de dados iríamos utilizar, se tínhamos algum algoritmo ou ferramenta em mente.

Informamos a base de dados, repassando a descrição disponível da base em sua página do kaggle, os objetivos informados foram os apresentados anteriormente, e não informamos nenhum algoritmo, buscando analisar qual seria sua sugestão.

Em resposta a esses dados, o ChatGPT apresentou uma sequência de passos7.2 para a produção desse trabalho, os quais podem ser divididos entre os passos da metodologia CRISP-DM e serão utilizados durante as próximas seções, e apresentou um código inicial**??** em que utiliza os métodos apriori e association rules da biblioteca mlxtend. Os passos apresentados foram:

1. Data Preparation
2. Exploratory Data Analysis (EDA)
3. Feature Selection
4. Association Rule Mining
5. Tools and Libraries
6. Interpretation and Reporting

2. Entendimento dos Dados

Nessa seção, será realizada uma análise da base de dados utilizada no artigo, seguindo as orientações do ChatGPT, e correspondendo ao passo de Exploratory Data Analysis citado anteriormente. A base de dados utilizada consiste em um conjunto de 6 datasets formado por playlists geradas automaticamente pelo spotify com músicas de gêneros semelhantes, sendo esses:

1. Alternativa com 25 playlists
2. Blues com 39 playlists
3. Hip Hop com 43 playlists
4. Indie com 55 playlists
5. Metal com 34 playlists
6. Pop com 63 playlists
7. Rock com 119 playlists

Todos esses datasets possuem a mesma estrutura com 22 colunas, estando entre elas:

1. Artist Name: The name of the artist who created the track.
2. Track Name: The name of the track.
3. Popularity: Popularity score of the track.
4. Genres: A list of genres associated with the track.
5. Playlist: The playlist to which the track belongs.
6. Various audio features such as danceability, energy, loudness, acousticness, and more.
7. Duration ms: The duration of the track in milliseconds.
8. Time signature: The time signature of the track.

Esses dados foram adquiridos sob a orientação do ChatGPT 7.3 e as descrições 7.4 de cada uma das colunas foi gerada por ele. Além disso, foi pedido por sugestões de escopos para a nossa análise considerando o dataset e as informações providas anteriormente, entre as sugestões estava regras de associação para playlists, o que coincide com o nosso objetivo inicial, dessa forma foi pedido para esse ser o foco da análise.

3. Preparação dos Dados

Em seguida ao entendimento dos dados que seriam utilizados foi necessária a preparação deles para a realização de futuras análises, correspondendo assim ao passo de Data Preparation sugerido pelo ChatGPT.

Primeiramente, perguntamos quais das colunas de nosso dataset seriam relevantes para a nossa análise e consequentemente deveriam ser mantidas no nosso dataframe, obtendo como resposta ”Artist Name”, ”Track Name”, ”Genres”e ”Playlist”. Em seguida, foram aplicados os passos que foram julgados relevantes entre os recomendados 7.5 por ele para o tratamento de dados, o que consistiu na remoção de valores nulos, uma vez que suas outras recomendações eram para métodos que saiam do escopo de nosso projeto.

Após esses passos, os dataframes analisados consistem de 4 colunas, com os atributos citados anteriormente e não foi necessário realizar a remoção de nenhuma das linhas devido a ausência de valores nulos.

4. Modelagem

Tendo em vista a limitação de páginas desse documento, decidiu-se que apenas o dataframe do rock seria analisado, já que ele é o que possui mais playlists. Além disso, pelo mesmo motivo, a mineração de cojuntos frequentes foi feito apenas para aritstas e músicas, mas não para subgêneros. Apesar dessas, limitações, a análise feita para esse dataframe pode ser facilmente a estrapolada para o restante dos gêneros.

Para computar os conjuntos frequentes, com a ajuda do chat GPT 7.6, criou-se uma função, também com ajuda da ferramenta automatizada, que recebe um dataframe, o nome de uma coluna e um suporte mínimo. Esse procedimento primeiramente codifica os dados de cada em um formato binário e posteriormente aplica o algoritmo a priori para encontrar os conjuntos frequentes. A biblioteca usada para essa função foi a mlxtend.

Para escolha do suporte mínimo, foi considerado que um valor muito alto poderia levar a regras muito gerais e muitos conjuntos frequentes de tamanho 1, já valores muito altos poderia levar a regras mais específicas. Então, a sugestão do chat foi iniciar com números mais próximo de 1, gerar as regras e avaliá-las, abaixar o suporte mínimo e seguir com esse processo iterativamente até chegar em um valor que gere regras interessantes com uma quantidade considerável de conjuntos frequentes. Dito isso, o valor escolhido para essa métrica foi de 0.07 para os conjuntos de artistas e 0.05 para as músicas, o que gerou 261 e 73 conjuntos, respectivamente.

Tabela 1. 10 artistas mais frequentes

suporte	itemsets
0.218487	(Foo Fighters)
0.184874	(Royal Blood)
0.176471	(Nirvana)
0.176471	(Pearl Jam)
0.176471	(U2)
0.159664	(Muse)
0.159664	(The Black Keys)
0.159664	(The Smashing Pumpkins)
0.159664	(Red Hot Chili Peppers)
0.151261	(Kings of Leon)

Tabela 2. 10 conjuntos de artistas mais frequentes com no mínimo 2 elementos

suporte	itemsets
0.134454	(Pearl Jam, Foo Fighters)
0.126050	(Nirvana, The Smashing Pumpkins)
0.117647	(Led Zeppelin, The Rolling Stones)
0.117647	(Kings of Leon, Foo Fighters)
0.117647	(Nirvana, Foo Fighters)
0.109244	(The Smashing Pumpkins, Foo Fighters)
0.109244	(Red Hot Chili Peppers, Foo Fighters)
0.109244	(Red Hot Chili Peppers, Nirvana)
0.100840	(Foo Fighters, Greta Van Fleet)
0.100840	(Nirvana, Stone Temple Pilots)

Tabela 3. 10 músicas mais frequentes

suporte	itemsets
0.109244	(When You Were Young)
0.084034	(Seven Nation Army)
0.084034	(hollywood sucks//)
0.075630	(Seize the Power)
0.075630	(Drive)
0.067227	(Dirty)
0.067227	(t r a n s p a r e n t s o u l feat. Travis Barker)
0.058824	(Again)
0.058824	(Hate To Say I Told You So)
0.058824	(Take Me Out)

Tabela 4. 10 conjuntos de músicas mais frequentes com no mínomo 2 alementos

suporte	itemsets
0.05042	(Seven Nation Army, Last Nite)
0.05042	(hollywood sucks//, When You Were Young)
0.05042	(Seven Nation Army, Are You Gonna Be My Girl, S...
0.05042	(hollywood sucks//, t r a n s p a r e n t s o u...
0.05042	(hollywood sucks//, love race (feat. Kellin Qui...
0.05042	(Seven Nation Army, Take Me Out)
0.05042	(Just Kidding, hollywood sucks//)
0.05042	(hollywood sucks//, Amnesia)
0.05042	(Seven Nation Army, Are You Gonna Be My Girl)
0.05042	(Are You Gonna Be My Girl, Steady, As She Goes)

5. Avaliação

Segundo o ChatGPT 7.7, para uma análise mais precisa de quais regras são de fato interessantes e trazem alguma informação, é preciso analisar mais que uma métrica. Então, foram escolhidas as métricas de suporte, confiança, lift e convicção.

A ferramenta automatizada também sugeriu um scatter plot 7.7 relacionando a confiança e suporte e outro para convicção e lift. Essa sugestão só foi acatada para o conjunto de artistas, visto que o conjunto de regras geradas no conjunto de músicas foi pequeno demais para traçar uma análise sobre esses gráficos.

5.1. Avaliação para o conjunto de artistas

Ao total foram geradas 119 regras para os conjuntos frequentes e foi escolhido um treshold de confiança 0.75 para as regras.

Tabela 5. As 10 regras com maiores lift

antecedentes	consequentes	suporte	confiança	lift	convicção
(Foo Fighters, Incubus)	(The Smashing Pumpkins, Kings of Leon)	0.07563	0.818182	10.818182	5.084034
(The Smashing Pumpkins, Kings of Leon)	(Foo Fighters, Incubus)	0.07563	1.000000	10.818182	inf
(Foo Fighters, Incubus)	(Nirvana, Kings of Leon)	0.07563	0.818182	9.736364	5.037815
(Nirvana, Kings of Leon)	(Foo Fighters, Incubus)	0.07563	0.900000	9.736364	9.075630
(Franz Ferdinand)	(The Raconteurs)	0.07563	0.900000	8.925000	8.991597
(grandson, KennyHoopla)	(MOD SUN)	0.07563	0.900000	8.925000	8.991597
(Nirvana, Incubus)	(The Smashing Pumpkins, Foo Fighters)	0.07563	0.900000	8.238462	8.907563
(The Smashing Pumpkins, Stone Temple Pilots)	(Nirvana, Pearl Jam)	0.07563	0.818182	8.113636	4.945378
(KennyHoopla, Machine Gun Kelly)	(MOD SUN)	0.07563	0.818182	8.113636	4.945378
(The Smashing Pumpkins, Kings of Leon, Foo Figh...	(Incubus)	0.07563	1.000000	7.933333	inf

Tabela 6. As 10 regras com maiores suporte

antecedentes	consequentes	suporte	confiança	lift	convicção
(Pearl Jam)	(Foo Fighters)	0.134454	0.761905	3.487179	3.282353
(The Smashing Pumpkins)	(Nirvana)	0.126050	0.789474	4.473684	3.911765
(Led Zeppelin)	(The Rolling Stones)	0.117647	0.875000	5.784722	6.789916
(The Rolling Stones)	(Led Zeppelin)	0.117647	0.777778	5.784722	3.894958
(Kings of Leon)	(Foo Fighters)	0.117647	0.777778	3.559829	3.516807
(Stone Temple Pilots)	(Nirvana)	0.100840	0.800000	4.533333	4.117647
(Stone Temple Pilots)	(Foo Fighters)	0.100840	0.800000	3.661538	3.907563
(MOD SUN)	(KennyHoopla)	0.100840	1.000000	7.437500	inf
(KennyHoopla)	(MOD SUN)	0.100840	0.750000	7.437500	3.596639
(The Smashing Pumpkins, Foo Fighters)	(Nirvana)	0.092437	0.846154	4.794872	5.352941

Por meio das tabelas, é possível extrair algumas informações interessantes. Nesse contexto, as regras com grande lift e convicção denotam uma forte relação entre os antecedentes e os consequentes daquela regra. Além disso, valores de confiança perto de 1 sugerem que se o antecedente da regra apareceu em uma playlist, a chance do consequente aparecer na mesma playlist é bem alta. Nesse raciocínio, a primeira regra da tabela 3 representa uma regra interessante em termos dessas métricas.

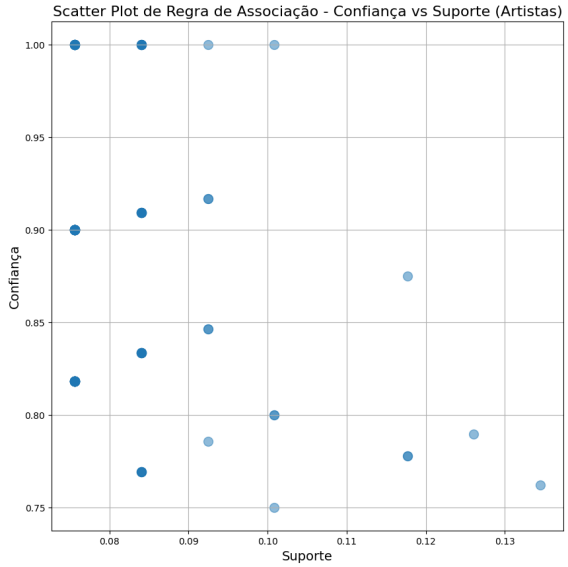

Figura 1. Relação entre suporte e confiança

O plot de confiança versus suporte sugere que há muitos dados com suporte de menor valor e uma quantidade considerável de regras com o mesmo valor de confiança. Isso pode ser explicado pela quantidade limitada de artistas no geral, o que faz que muitos subconjuntos fiquem com a mesma frequência. Além disso, o gráfico também parece sugerir que quanto maior o suporte de uma regra, há menos chance de ela ter uma alta confiança.

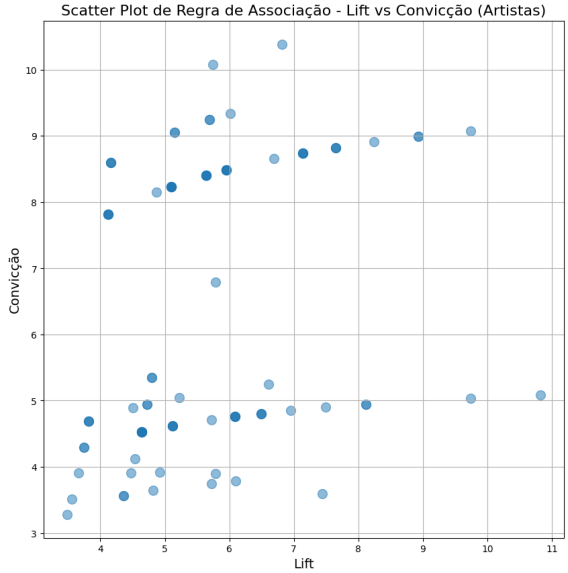

Figura 2. Relação entre lift e convicção

Os valores nesse gráfico parecem estar concentradas em duas faixas, o que também pode ser explicado pelo fato de vários conjuntos possuírem o mesmo suporte. Ademais, o parece haver uma correlação positiva entre lift e convicção.

5.2. Avaliação para o conjunto de músicas

Ao total, foram geradas 26 regras para os conjuntos frequentes e foi escolhido um treshold de confiança 0.75 para as regras.

Tabela 7. As 10 regras com maiores lift

antecedentes	consequentes	suporte	confiança	lift	convicção
(Steady, As She Goes)	(Are You Gonna Be My Girl)	0.05042	1.000000	19.833333	inf
(Are You Gonna Be My Girl)	(Steady, As She Goes)	0.05042	1.000000	19.833333	inf
(Seven Nation Army, Steady, As She Goes)	(Are You Gonna Be My Girl)	0.05042	1.000000	19.833333	inf
(Are You Gonna Be My Girl)	(Seven Nation Army, Steady, As She Goes)	0.05042	1.000000	19.833333	inf
(Steady, As She Goes)	(Seven Nation Army, Are You Gonna Be My Girl)	0.05042	1.000000	19.833333	inf
(Seven Nation Army, Are You Gonna Be My Girl)	(Steady, As She Goes)	0.05042	1.000000	19.833333	inf
(Hate To Say I Told You So)	(Seven Nation Army, Take Me Out)	0.05042	0.857143	17.000000	6.647059
(Seven Nation Army, Take Me Out)	(Hate To Say I Told You So)	0.05042	1.000000	17.000000	inf
(Seven Nation Army, Hate To Say I Told You So)	(Take Me Out)	0.05042	1.000000	17.000000	inf
(Take Me Out)	(Seven Nation Army, Hate To Say I Told You So)	0.05042	0.857143	17.000000	6.647059

Tabela 8. As 10 regras com maiores suporte

antecedentes	consequentes	suporte	confiança	lift	convicção
(Last Nite)	(Seven Nation Army)	0.05042	1.000000	11.900000	inf
(Seven Nation Army, Are You Gonna Be My Girl)	(Steady, As She Goes)	0.05042	1.000000	19.833333	inf
(Hate To Say I Told You So)	(Seven Nation Army, Take Me Out)	0.05042	0.857143	17.000000	6.647059
(Seven Nation Army, Take Me Out)	(Hate To Say I Told You So)	0.05042	1.000000	17.000000	inf
(Take Me Out, Hate To Say I Told You So)	(Seven Nation Army)	0.05042	1.000000	11.900000	inf
(Seven Nation Army, Hate To Say I Told You So)	(Take Me Out)	0.05042	1.000000	17.000000	inf
(Take Me Out)	(Hate To Say I Told You So)	0.05042	0.857143	14.571429	6.588235
(Hate To Say I Told You So)	(Take Me Out)	0.05042	0.857143	14.571429	6.588235
(No One Knows)	(Seven Nation Army)	0.05042	1.000000	11.900000	inf
(SOS (feat. Travis Barker))	(hollywood sucks//)	0.05042	1.000000	11.900000	inf

Em comparação com o conjunto de artistas, as regras de músicas tem menor suporte, o que indica que há conjuntos de músicas que se repetem em várias playlists do que conjuntos de artistas ou bandas que se repetem em várias playlists. Isso faz sentido porque um único músico pode ter várias faixas em várias playlists, o que aumenta a chance dos conjuntos ocorrerem mais que uma vez. No entanto, há regras interessantes, por exemplo, playlists que contêm a música Last Nite, irão, com certeza, conter a música Seven Nation Army.

6. Conclusões e perspectivas

Em relação ao ChatGPT, o seu desempenho variou com base em cada uma das etapas do processo de construção do artigo, e será analisado para cada um dos passos do CRISP-DM, de forma a corresponder a cada seção desse trabalho.

6.1. Entendimento do Negócio

Na fase de Entendimento do Negócio foi possível perceber a capacidade da ferramenta de direcionar um projeto em que o utilizador dela não compartilha as informações necessárias e consequentemente a ferramenta precisa requisitá-las.

Destaca-se a criação de um passo a passo pelo ChatGPT com o objetivo de facilitar o processo da criação do código e análise dos dados. Entretanto, é importante ressaltar

que cada um dos passos é apresentado de forma muito resumida e alguns aparentam ser desnecessários, como o passo Tools and Libraries, que consistia somente na inclusão de uma biblioteca.

6.2. Entendimento dos Dados

Durante o Entendimento dos Dados a ferramenta se mostra novamente útil ao ser capaz de utilizando a saída de códigos que foram disponibilizados por ela, inferir uma descrição de cada um dos atributos do dataset. E, além disso, foi capaz de gerar sugestões relevantes de escopos para o projeto, com diferentes complexidades e áreas.

6.3. Preparação dos Dados

Na Preparação dos Dados, foi possível perceber uma das imperfeições do ChatGPT, a análise do banco de dados foi muito superficial e o objetivo definido anteriormente foi ignorado, em consequência disso, a maior parte das sugestões de métodos de preparação de dados era desnecessária e não se aplicava ao escopo definido.

Esse é um padrão que pode ser percebido durante o uso do ChatGPT, sempre há a necessidade de haver uma resposta, mesmo quando a ferramenta não tem certeza sobre a sua veracidade ou relevância, o que pode ser muito prejudicial para utilizadores ao perguntarem sobre áreas em que são leigos, possivelmente sendo uma fonte de desinformação.

6.4. Modelagem

A ferramenta mostrou-se especialmente útil na modelagem dos dados. Com apenas algumas informações sobre a estrutura da tabela e o objetivo do trabalho, o ChatGPT foi capaz de criar uma função para computar os conjuntos frequentes de maneira assertiva. No entanto, é preciso ressaltar que é preciso ter cautela ao usar os pedaços de código oferecidos dessa maneira, visto que eles nem sempre estão certos. Então, para esse trabalho, procurou-se saber um pouco sobre com o a biblioteca utilizada na função e também foi pedida informações adicionais sobre o que procedimento faz de fato.

6.5. Avaliação

De forma análoga à modelagem, o chat também ofereceu pedaços de código que ajudaram na criação das regras e de novo uma pesquisa foi necessária para entender as funcionalidades da biblioteca utilizada. Além disso, ele sugeriu os plots usados na seção 5.1 e deu o código para gerá-los. Por fim, a ferramenta atuou como uma mentor ao tirar dúvidas sobre as métricas das regras, como lift, convicção e confiança 7.7.

Em suma, por meio desse trabalho, ficou evidente que o ChatGPT é uma ferramenta muito poderosa para a mineração de conjunto frequentes, mas é preciso ter cautela na hora utilizá-lo, pois ele tem suas limitações e às vezes comete erros.

7. Anexos

7.1. Primeira Interação

Primeira interação com o ChatGPT

7.2. Passo-a-Passo

Passo-a-passo para um trabalho de Mineração de dados

Passo-a-passo para um trabalho de Mineração de dados (segunda parte)

7.3. Extração de Dados

Instruções do ChatGPT para extração de informações do dataset

7.4. Entendimento dos Dados

Interpretação dos dados do dataset pelo ChatGPT

7.5. Preparação dos Dados

Instruções para preparação dos dados

Instruções para preparação dos dados (segunda parte)

7.6. Modelagem

Função para computar conjuntos frequentes

7.7. Avaliação

Função para computar regras de associação

Visualização das regras

Explicação das Regras

7.8. Notebook

Link para notebook

Referências

Zaki, M. J.,& Meira, W., Jr. (2020). Data mining and machine learning: Fundamental concepts and algorithms (2o ed). Cambridge University Press.

([S.d.]). Openai.com. Recuperado 1o de outubro de 2023, de https://chat.openai.com/