iscte INSTITUTO UNIVERSITÁRIO DE LISBOA

MUSIFY AGENCY

PCA & Clustering

Licenciatura em Ciência de Dados

Métodos de Aprendizagem Não Supervisionada

Docente: José Dias

Ana Lapa, nº 93111
João Ribeiro, nº 93067
Ricardo Frazão, nº 92630
Ricardo Mendes, nº 92564
Rita Bairros, nº 92692

04/06/2021

Índice

Índice de Tabelas

1. Introdução

O Spotify é um serviço de streaming digital que dá acesso instantâneo a milhões de músicas, podcasts, entre outros conteúdos de diversos artistas de todo o mundo. Nos últimos anos, a quantidade de utilizadores desta plataforma, tem apresentado um grande crescimento. Este aumento pode ser explicado pela facilidade de acesso a este serviço nos mais variados dispositivos.

Assim, torna-se bastante relevante analisar as preferências musicais dos utilizadores do Spotify. Deste modo, utilizar-se-á uma base de dados que contêm informação relativa ao Top das músicas mais ouvidas, por cada ano, durante o período de 2010 a 2019.

Tendo em conta o reconhecimento desta aplicação, a informação presente neste dataset poderá ser útil a qualquer pessoa que esteja envolvida na indústria musical, de modo a potenciar o seu sucesso.

A Musify Agency é uma nova empresa que representa artistas no ramo da música. Esta, ao tentar destacar-se neste mercado, pretende obter um maior conhecimento relativamente às características ideais dos artistas em que deve apostar.

Com base nesta informação, esta agência solicitou um estudo que tem como objetivo analisar os géneros de música mais procurados, com vista a representar artistas cujo género musical coincida com a preferência dos utilizadores. Para além disto, a empresa tenciona compreender se o sexo do artista é um fator decisivo na seleção das músicas por parte do público. Por fim, a Musify também tem interesse em verificar qual o tipo de artistas (a solo, duplas ou bandas), cujas músicas apresentam uma maior popularidade.

Com vista a alcançar o objetivo anteriormente mencionado, este relatório irá incluir todos os procedimentos necessários para o atingir. Deste modo, será retratada a descrição da base de dados e o seu respetivo tratamento. Posteriormente, proceder-se-á à redução da dimensão das variáveis, seguida da realização de um modelo de clustering. Para finalizar, apresentar-se-á uma conclusão que irá resumir todo o conteúdo abordado ao longo do relatório.

2. Descrição dos dados

O dataset que nos fora disponibilizado através do website Kaggle, contém dados acerca do Top de músicas mais ouvidas em cada ano, entre 2010 e 2019, na plataforma Spotify. Este apresenta 603 observações e 14 variáveis, das quais 3 são categóricas e 11 são numéricas.

As variáveis utilizadas na análise das componentes principais (PCA) são consideradas como variáveis ativas (Input), sendo neste caso, a variável *Beats Per Minute, Energy, Danceability, Loudness, Liveness, Valence, Acousticness, Speechiness, Popularity,* e *Duration*. Relativamente às variáveis passivas (Profile), estas são utilizadas para a caracterização dos clusters, tendo sido utilizada a variável *Genre*, *Gender* e *Type*. Estas duas últimas foram manualmente adicionadas, com o objetivo de acrescentar informação ao dataset.

Assim, a partir da realização do clustering iremos obter mais conhecimento à cerca de cada grupo, o que facilitará a tomada de decisão por parte da Musify Agency.

3. Tratamento do Dataset

Uma vez conhecida a estrutura do dataset, procedeu-se à análise das medidas descritivas do mesmo. Através deste processo, foi possível verificar que a base de dados não apresentava valores omissos. Para além disto, considerou-se pertinente efetuar a conversão da variável *year* de numérica para categórica, visto que é mais interessante analisar o seu desempenho, estando esta representada por categorias.

Com o intuito de obter mais informação relativamente às variáveis ativas, procedeu-se à visualização da correlação entre estas. Adicionalmente, ao analisar os gráficos, foi possível verificar que as variáveis *Beats Per Minute* (*bpm*) e *Loudness* (*dB*) apresentavam outliers. Esta afirmação é sustentada pelo facto de a variável *bpm* não poder conter valores nulos e a variável dB apresentar um valor que se distancia em muito dos restantes. Ao analisar estes valores, observou-se que ambos pertenciam à mesma música, tendo-se procedido à remoção da mesma.

Posteriormente, de modo a facilitar a análise da variável *genre* na interpretação do clustering, foram alteradas o número de categorias desta. Assim, em vez de a variável apresentar as 50 classes iniciais passará a conter apenas 12 categorias mais gerais.

4. Identificação das dimensões da análise

Ao analisar as características das variáveis foi possível verificar que era relevante efetuar a análise das componentes principais (PCA). Tendo em consideração as 10 variáveis Input utilizadas, optou-se pelo processo de redução de dimensão destas.

No processo de análise e interpretação das PC's, foram, em primeiro lugar, criadas 4 componentes principais. Contudo, esta opção foi descartada, visto que a variância explicada seria cerca de 9% inferior, em relação à utilização de 5 componentes principais. Dado isto, a interpretação resultante destas é bastante mais intuitiva e fundamentada.

Assim, com base nesta informação, decidiu-se utilizar 5 componentes principais, sendo que estas explicam cerca de 71,1% das variáveis originais. Ao fazê-lo, procedeu-se à interpretação de cada componente principal, tendo-se atribuído um nome representativo das suas características predominantes.

Relativamente à primeira componente principal, esta é evidenciada por um elevado volume e energia. Por outro lado, apresenta uma acústica reduzida, sendo por isso denominada por "*eletronic*". Seguindo o mesmo critério, à segunda componente principal (PC2) foi atribuída a designação "*sad_long_songs*" devido à elevada presença da variável *duration* e ao impacto negativo da variável *valence*. Atendendo à clara influência da variável que expressa o número de palavras na música, associa-se à PC3 o nome "*rap*". Tendo em consideração que as variáveis *bpm* e *dance* são dominantes de forma negativa e positiva, respetivamente, na quarta componente principal, a esta foi-lhe atribuída o nome *"slows"*. Por fim, a variável *pop* evidencia o maior loading de toda a análise, sendo *"popular"* a designação que melhor caracteriza a quinta componente principal.

Nome	Descrição
PC1 - Eletronic	Energia e Volume - Aumenta Acústica - Diminui
PC2 - Sad_long_songs	Duração e o facto de ser ao vivo - Aumenta Positividade - Diminui
PC3 - Rap	Quantidade de Palavras - Aumenta
PC4 - Slows	Vontade de dançar - Aumenta Ritmo da música - Diminui
PC5 - Popular	Popularidade da música - Aumenta

Tabela 1 - Componentes principais e respetivas características

5. Identificação da heterogeneidade na base de dados

Uma vez definidas as componentes principais a utilizar, torna-se bastante relevante proceder a uma análise baseada em clustering. Com vista a fornecer à Musify Agency as informações por esta solicitadas, recorreram-se a diversas técnicas da aprendizagem não supervisionada para o efeito. Estas incluem algoritmos como o Clustering Hierárquico, o K-means, o PAM e o GMM (Clustering Probabilístico). Assim, para dar resposta ao problema, optou-se por recorrer ao algoritmo K-means.

Posto isto, tendo por base as variáveis Profile anteriormente definidas, a cada um dos clusters criados foi atribuído um nome capaz de ilustrar as suas características.

No que diz respeito ao primeiro cluster, este é caracterizado por uma elevada presença de artistas a solo (80%), em relação à variável *type*. Contudo, nesta variável é possível observar que a categoria *duo*, apresenta a maior parte dos seus artistas neste cluster. Para além disto, verifica-se uma superioridade de 13 pontos percentuais do sexo masculino em comparação com o feminino, na variável *gender*. Relativamente à variável *genre* averigua-se que o género de música mais predominante é o *pop*, representando 40% deste cluster. Dado isto, a designação atribuída a este é "*Male_solo_duo_pop*".

Prosseguindo para o cluster seguinte, este assemelha-se ao anterior, na variável *gender*. O género musical com maior influência neste cluster é o *dance pop* (60%). Adicionalmente, o tipo de artista que apresenta um maior destaque são os artistas a solo. Todavia, é de referir o facto da classe *band* influenciar em cerca de 18% a variável *type* neste cluster. Deste modo, "*Male_solo_band_dance*" é o nome que melhor caracteriza este cluster.

Em relação ao terceiro cluster observa-se uma predominância significativa da categoria *dance pop*, correspondendo a 73% dos géneros musicais que neste figuram. No entanto, existem outras duas categorias mais evidenciadas entre as restantes (*pop* e *hip hop*). À semelhança dos clusters acima mencionados, a classe *solo* evidencia uma percentagem consideravelmente superior em comparação com as restantes. Ainda que seja notória uma elevada presença de artistas do sexo masculino, é possível destacar a grande quantidade de artistas femininos. Assim, este cluster fora denominado por "*Male_solo _dance*".

Por fim, o último cluster apresenta uma percentagem quase idêntica de artistas de ambos os sexos, sendo o sexo feminino um pouco superior. Relativamente às variáveis *genre e type* estas são bastante semelhantes com o cluster 2. Com base nesta informação, "*Female_solo_dance*" é a melhor designação capaz de diferenciar este cluster.

	Male_solo_duo_pop	*Male_solo_band_dance*	*Male_solo_dance*	*Female_solo_dance*
Genre	40% *pop* 32% *dance pop*	60% *dance pop* 26% *pop*	73% *dance pop* 15% *pop* 8% *hip hop*	42% *dance pop* 35% *pop*
Gender	41% *Female* 54% *Male* 4% *both*	45% *Female* 52% *Male* 3% *both*	45% *Female* 48% *Male* 7% *both*	49% *Female* 48% *Male* 4% *both*
Type	80% *solo* 12% *band* 8% *duo*	79% *solo* 18% *band* 3% *duo*	82% *solo* 11% *band* 7% *duo*	83% *solo* 14% *band* 2% *duo*

Tabela 2- Caracterização dos clusters com base nas variáveis profile

Em suma, comparando todos os clusters, é notável a dominância clara da classe *dance pop* em, dado que 55,3% do dataset é composto por esta. Esta afirmação pode ser explicada pelo facto de esta representar a maior parte do dataset nesta variável. Relativamente à coluna *gender*, o primeiro cluster apresenta a maior discrepância de valores entre os dois sexos (13%). Todavia, ao considerar o dataset num todo, é possível verificar que ambos os sexos apresentam uma percentagem semelhante na base de dados. Ainda assim, é de frisar o facto de a quantidade de indivíduos do sexo masculino exceder em 6% o número de indivíduos do sexo feminino. Por fim, em relação à variável *type* é possível compreender que a categoria *solo* lidera todos os clusters, uma vez que representa 80,2% da base de dados. Apesar disto, a segunda categoria mais evidenciada nesta variável é a categoria *band*, sendo esta sobressaída no segundo cluster (18%).

6. Conclusão

Tendo em conta a popularidade associada à plataforma Spotify, recorreu-se a uma base de dados que contem dados relativos ao Top de músicas mais ouvidas. Tendo esta como referência, tornou-se pertinente efetuar a análise da mesma, com vista a adquirir um maior conhecimento acerca desta área.

Deste modo, a pedido da Musify Agency, elaborou-se este estudo com vista a analisar qual o tipo de artistas mais recomendado a representar. Através deste será possível verificar quais os melhores artistas a investir com vista a maximizar o lucro desta agência.

Relativamente às músicas mais reconhecidas na plataforma Spotify, o género musical que apresenta uma maior popularidade é, destacadamente, o *dance pop*. A par deste, pode-se ainda salientar o género musical *pop*, como o segundo mais procurado pelos utilizadores desta plataforma.

No que diz respeito ao tipo de artistas (*solo*, *duo* e *band*), verificou-se que a grande maioria das músicas mais ouvidas pertencem a artistas a solo. Esta informação é corroborada pela evidente dominância da classe *solo* em todos os clusters criados.

Com base na análise realizada anteriormente, constatou-se que o sexo, representado pela variável *gender*, não é um fator determinante na escolha das músicas preferidas dos utilizadores. Assim, é possível aferir que, na indústria musical, existe uma igualdade de oportunidades para ambos os sexos.

Em suma, sugere-se à Musify Agency a contratação de artistas que produzam músicas a solo na vertente *dance pop*, independentemente do género do artista.

7. Referências Bibliográficas

1. Kaggle. 2019. *Top Spotify songs from 2010-2019 – BY YEAR.* Consultado em 4 de maio 2021. Disponível em https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year
2. Kaggle. 2021. *Spotify 2010-2019 Insight Analysis with R*. Consultado em 20 de maio 2021. Disponível em https://www.kaggle.com/trixietacung/spotify-2010-2019-insight-analysis-with-r