Receita para o Sucesso - Uma análise das características que são comuns às músicas que se tornaram populares no TikTok

Daniel Barreto Torres[1], **Gabriela Tavares Barreto**[1], **Guilherme Lucas Giudice Silva**[1], **Mirna Mendonça e Silva**[1] e **Vinicius Silva Gomes**[1]

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

{daniel.torres, gabrielabarreto, guilhermegiudice, mirnamendonca, vinicius.gomes}@dcc.ufmg.br

Resumo. *Este artigo discute um processo de manipulação e análise de dados de músicas que fizeram sucesso na rede social TikTok. O objetivo central do estudo é tentar identificar se existe algum conjunto de características comum a essas músicas, de modo que seja possível identificar esse grupo e, então, avaliar se é possível formular uma "receita" para o sucesso de uma música no aplicativo.*

1. Introdução

Com mais de 1 bilhão de usuários ativos por mês em 2021, o TikTok se tornou uma das redes sociais mais usadas no mundo inteiro, junto com outras redes gigantes, como Facebook e Instagram, por exemplo [Carolina Walliter 2021]. Diferentemente dessas duas redes sociais, que permitem o compartilhamento de fotos e vídeos, o TikTok é voltado exclusivamente para a interação entre os usuários através da publicação de vídeos.

Esses vídeos que circulam na plataforma são gravados e compartilhados pelos próprios usuários e possuem os mais variados objetivos, indo desde o entretenimento até o compartilhamento de conteúdos educativos. Por esse motivo, a rede social atraiu, principalmente, o público jovem (entre os 16 e 24 anos), sendo este grupo responsável por 66% do total de usuários da rede [Carolina Walliter 2021].

Se popularizando durante o cenário de pandemia que assolou o globo em 2020 e 2021, o TikTok ganhou forças como uma rede social onde as pessoas poderiam se entreter durante o isolamento social e, ainda, produzir seus próprios vídeos para interagir com os outros usuários.

Nesse sentido, alguns tipos de vídeos e conteúdos se popularizaram mais e ganharam mais destaque que os outros, ganhando referência em diversas outras situações, como propagandas de televisão, comemorações de futebol, etc.

No entanto, é possível fazer a seguinte pergunta: o que aconteceu para que esses conteúdos especificamente se popularizassem mais que os outros? Quais características esses vídeos possuem que fizeram com que eles se destacassem mais perante a uma quantidade gigantesca de vídeos que eram publicados diariamente no aplicativo.

Partindo dessas dúvidas, um estudo foi desenvolvido para tentar identificar quais eram essas características que faziam com que algo se destacasse mais no TikTok. Entretanto, o escopo principal do projeto será identificar as características das músicas que mais fizeram sucesso no aplicativo, tentando observar se elas possuem algo em comum ou se existe algum outro tipo de atributo compartilhado por elas.

2. Objetivos

Assim como dito na seção anterior, partindo da percepção que alguns conteúdos ganham mais destaque na plataforma do que outros, algumas músicas também ficaram mais famosas do que outras na plataforma, os *hits* mundialmente conhecidos.

Mais que isso, muitas músicas mais antigas e que já não fazem tanto sucesso tiveram picos de audiência em vários serviços de *streaming* após a aparição dessas músicas em vídeos que fizeram muito sucesso no aplicativo. Por ficarem muito conhecidas no aplicativo, consequentemente, os artistas que produziram essas músicas também ficam. Sendo assim, muitos artistas passaram a compor músicas de modo que o seu sucesso na plataforma seja facilitado.

Dito isso, o objetivo principal deste estudo é tentar identificar se existem, de fato, características que são comuns as músicas que fazem sucesso na plataforma e se existe mesmo alguma "receita" que favoreça isso.

Além disso, pretendemos observar, de maneira intermediária, se existem relações entre as características entre si e se existe algum motivo para que uma música muito antiga seja revivida na rede social e volte a fazer muito sucesso mundialmente.

3. Datasets e dados coletados

Para fazer as análises que desejamos, escolhemos para trabalhar um dataset que contém diversos atributos das músicas mais populares do TikTok nos últimos anos. Se trata de um dataset público, encontrado na comunidade de ciência de dados e aprendizado de máquina Kaggle [Kaggle 2022].

O dataset se chama "TikTok Trending Tracks" e contém atributos bem importantes para as pesquisas que pretendemos fazer, como a popularidade da música, a dançabilidade, o gênero, atributos relacionados com a parte instrumental, tom, entre outras.

As informações que compõem o dataset são provenientes da API pública do Spotify [Spotify 2022]. Dessa forma, atributos que não são originados de contagens ou outras formas de coleta comuns, como dançabilidade e popularidade, são métricas calculadas pela própria equipe do Spotify. Além disso, o autor do dataset não especifica o processo de escolha das músicas que o compõem.

No entanto, após uma análise feita pelos membros do grupo que realizam o estudo, e que também são usuários da rede social, foi possível constatar que as músicas escolhidas fazem sentido e são músicas que o grupo, como usuários, perceberam sua presença enquanto utilizavam o aplicativo.

Além disso, um processo de *web scrapping* foi realizado para enriquecer o estudo e auxiliar no processo de clusterização, para que as conclusões tiradas até aquele momento sejam melhor embasadas e clareadas. Esse processo será melhor explicado na seção 4, onde a análise exploratória sobre os dados será discutida.

4. Manipulação e análise do dataset

Nessa seção será melhor destrinchado o processo de manipulação e análise dos dados escolhidos para avaliar o sucesso das músicas no TikTok. O primeiro passo realizado foi

o *download* do dataset principal e a coleta de dois outros grupos externos de músicas para enriquecer a nossa análise ao final.

Após isso, os dados foram tratados para que a manipulação e a interpretação seja facilitada. Em seguida, o processo de análise exploratória começou, onde as primeiras características dos dados puderam ser observadas.

Seguido dele, houve a tentativa de algumas regressões e uma clusterização, onde finalmente pudemos avaliar as nossas hipóteses e interpretar se, de fato, existe alguma relação entre os atributos das músicas e o sucesso delas no aplicativo. As subseções a seguir descrevem melhor cada etapa e apresentam os resultados parciais obtidos.

Todos os processos que serão discutidos foram realizados utilizando a linguagem de programação **Python**, além de pacotes relacionados a ciência de dados bastante conhecidos e que podem ser importados pela linguagem. O repositório no GitHub[1] armazena todos os códigos-fonte e gráficos produzidos nesse estudo.

4.1. Tratamento dos dados

O primeiro passo no tratamento dos dados foi investigar qual era o formato do dataset escolhido. Inicialmente ele possuía 6746 músicas e 24 colunas de atributos.

Após isso, foi verificado que não havia nenhuma observação com dados faltantes mas haviam algumas linhas com dados duplicados. Com a remoção dos dados duplicados, o dataset passa a ter 4263 músicas.

Ao final, a última mudança foi a adição de uma coluna extra somente com o ano em que a música foi lançada, para facilitar a manipulação, uma vez que esse dado estava implicitamente colocado num campo com a data completa de lançamento.

Por fim, o dataset foi salvo em disco, juntamente com uma versão dele que possuía apenas os campos numéricos do dataset, para facilitar o uso para nas próximas etapas.

4.2. Análise exploratória

O primeiro trabalho de análise consistiu em calcular as medidas de tendência central para cada atributo presente no dataset. Assim é possível avaliar seus comportamentos, tirar as primeiras conclusões, e clarear o direcionamento dos *insights* relacionados ao conjunto de dados.

Em seguida, o histograma de alguns atributos foi computado para observar a distribuição dos valores para cada característica. Com isso, através de gráficos como os das figuras 1 e 2, foi possível concluir que a popularidade das músicas do dataset no Spotify varia de acordo com uma distribuição que se parece com uma bimodal, onde um dos picos é próximo do 0, mostrando que existem músicas que fizeram sucesso no aplicativo mas que não fazem sucesso no Spotify.

Além disso, foi possível atestar pelo gráfico da figura 2 que as músicas que compõem o dataset são músicas majoritariamente recentes, lançadas em 2020 e 2021. Apesar disso, há a presença de músicas mais antigas na base, sendo a mais antiga lançada em 1960.

[1]Esse repositório está disponível e pode ser acessado através do link: `https://github.com/MirnaMendonca/TP-ICD`.

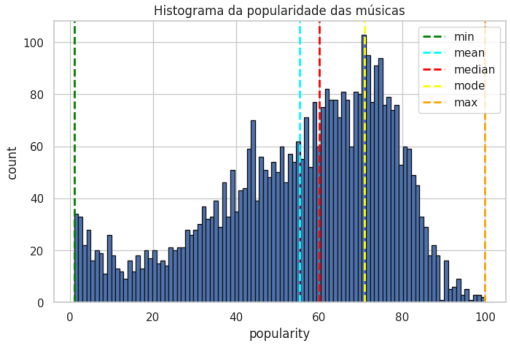

Figura 1. Histograma da popularidade das músicas presentes no dataset.

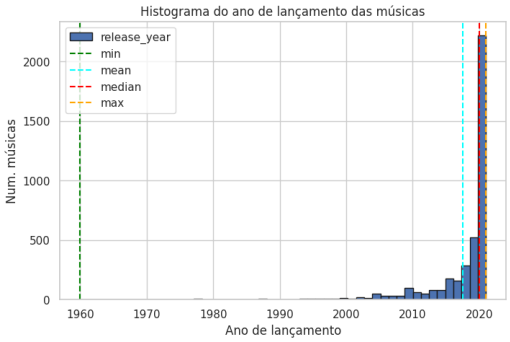

Figura 2. Histograma do ano de lançamento das músicas presentes no dataset.

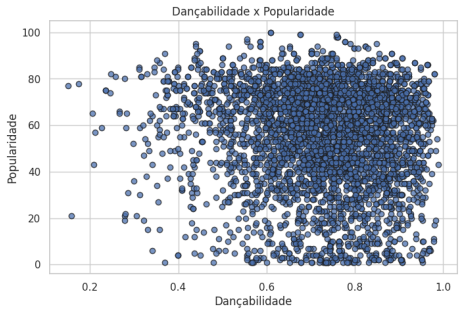

Figura 3. Gráfico de dispersão da dançabilidade das músicas pela popularidade.

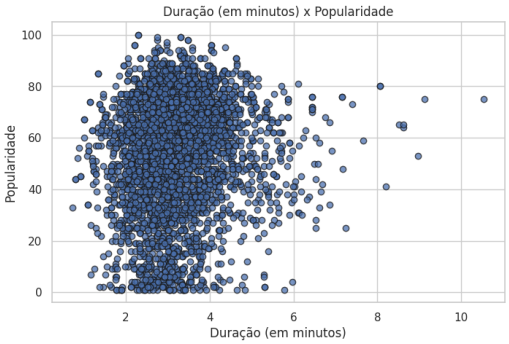

Figura 4. Gráfico de dispersão da duração das músicas pela popularidade.

Ademais, as relações entre os atributos foram estudadas para verificar se existe alguma correlação entre eles que possa ser explicada pela disposição dos pontos dois-a-dois. Nessa etapa, alguns gráficos, como os presentes nas figuras 3 e 4, foram construídos, para que a disposição dos pontos possa ser estudada.

Nesses gráficos foi possível concluir que não existe nenhuma relação muito clara entre a dançabilidade e a popularidade das músicas presentes no dataset. Em contrapartida, foi possível observar que existe uma concentração maior de músicas com duração entre 2 à 4 minutos.

Essa situação leva aos primeiros *insights* e possíveis questões futuras a serem realizadas, sobre o motivo para tal resultado, uma vez que a plataforma parece conter uma quantidade maior de músicas relativamente curtas.

4.3. Regressão

Para tentar enriquecer a análise, algumas regressões foram traçadas para identificar se o comportamento de alguns atributos pode ser capturado por modelos lineares.

A figura 5 mostra o *pair plot* de alguns dos principais atributos da base de dados. Esse *plot* contém vários gráficos de dispersão desses atributos juntamente com tentativas de regressões lineares para cada *plot*.

Como é possível observar, nenhum gráfico apresenta uma regressão que explica bem os dados. Sendo assim, parece que, caso os dados possam ser explicados por algum modelo, esse modelo deve ser, definitivamente, mais complexo que o linear.

Por fim, a figura 6 apresenta o *heatmap* das correlações presentes nos atributos. Assim como as regressões, esse gráfico apenas reforça a ideia de que o modelo linear não é bom para as observações contidas no dataset.

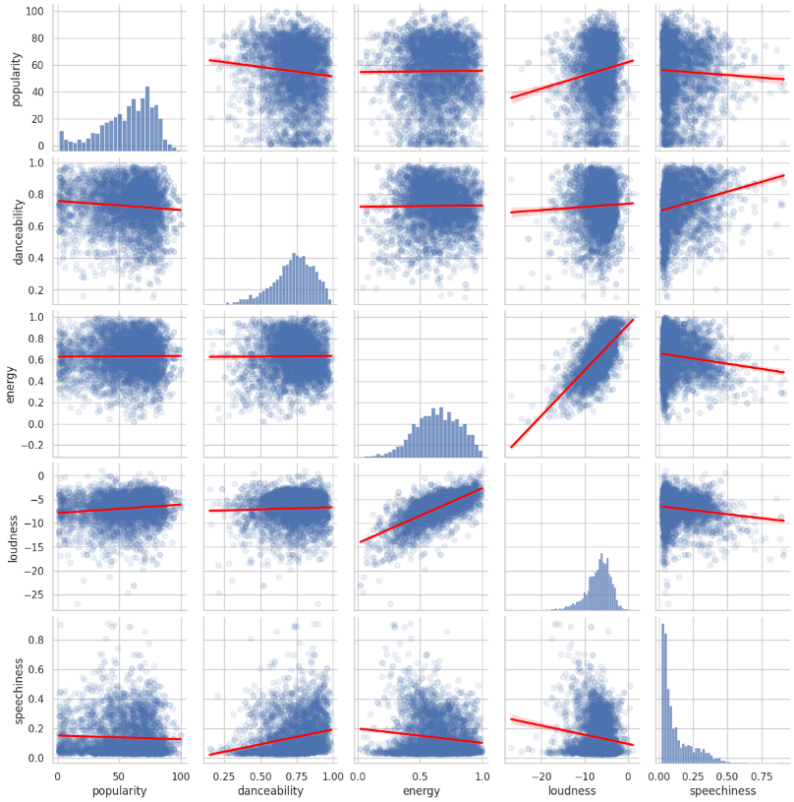

Figura 5. *Pair plot* dos atributos numéricos presentes no dataset.

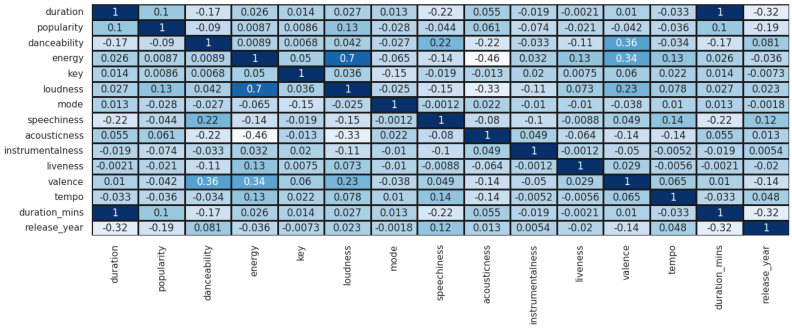

Figura 6. *Heatmap* da correlação dos atributos numéricos do dataset.

4.4. Análise de componentes principais - PCA

Neste trabalho, foi utilizado o método de análise via componentes principais de forma aprofundada em dois momentos, primeiro para tentar identificar subgrupos dentro das

músicas populares no TikTok, e em um segundo momento para poder comparar a base de músicas populares do TikTok com outras duas bases de músicas. Essas análises serão detalhadas a seguir.

4.4.1. Analisando as músicas populares no TikTok

A percepção dos integrantes do grupo, vinda do uso individual do TikTok, é que há muitos conteúdos diferentes disponíveis nessa rede, por exemplo, há conteúdo sobre política, animais, vídeos de maquiagem, danças e tudo mais que se possa imaginar. Ao mesmo tempo que essa pluralidade existe, também percebe-se a repetibilidade de determinadas *trends* (tendências, em português). Constantemente vemos alguma *trend* de dança fazendo sucesso, ou alguma *trend* de homenagem (*trends* em que filhos homenageiam os pais, ou namorados homenageiam as namoradas, entre outros), *trends* de maquiadores, *trends* relacionadas a algum lançamento recente de série ou filme, por exemplo. Isso é só para citar algumas. Muitas dessas *trends* envolvem músicas, sejam elas a parte principal do vídeo, como nas danças, ou de forma secundária, como nos vídeos de homenagem.

Isso levou ao seguinte questionamento: se não poderia haver então subgrupos distintos dentre as músicas populares no TikTok, que pudessem se dividir de acordo com o tipo de trend/nicho ao qual estavam associados. Por isso, foi optado em analisar via PCA essas músicas. O PCA foi escolhido por simplificar a análise via redução de dimensionalidade, facilitando encontrar as características distintivas entre esses grupos.

Primeiro, foi feita uma normalização da base de músicas, para depois aplicar o algoritmo de PCA disponibilizado pela biblioteca *SKLearn*. Tendo aplicado o PCA, foi identificado que todas componentes principais capturavam um percentual pequeno de variância da amostra, sendo que a primeira componente capturava somente 19%.

A partir desse resultado, notou-se que era possível reduzir a dimensão de elementos analisados para cinco componentes principais, de forma a capturar somente 60% da variância. Esse resultado já não era dos mais positivos, pois resultados ótimos via PCA conseguem obter uma redução de dimensionalidade para até duas componentes principais, explicando mais que 80% da variância.

Assim, dando prosseguimento a análise, foi gerado o *plot* dos dados transformados pela redução de dimensionalidade, que pode ser verificado na 7 para ter uma visualização do resultado obtido. Nesse *plot*, as músicas foram diferidas por popularidade.

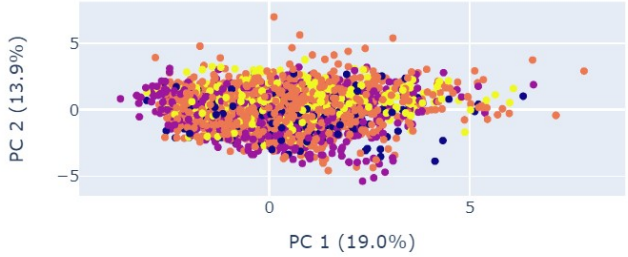

Figura 7. Gráfico de dispersão dos dados de acordo com a primeira e segunda componentes principais.

Em ordem crescente de popularidade temos as cores azul, roxo, laranja e amarelo. O resultado dessa análise apontam para diferentes interpretações. A primeira é que pelo menos para as músicas da base escolhida, não há muita diferenças entre as características das músicas das diferentes *trends*. Porém, como é desconhecida a forma que esses dados foram capturados, é possível criar a hipótese de que as músicas escolhidas estavam dentro de um nicho comum, e por isso diferem pouco entre si. Por último, esse resultado pouco conclusivo pode ser fruto da própria escolha do PCA como algoritmo para tentar explicar os dados, e é possível que outro tipo de análise chegasse a resultados mais conclusivos.

4.4.2. Analisando todas as músicas

A segunda tentativa de uso do PCA consistiu em analisar a junção de três diferentes datasets. O de músicas populares no TikTok, proveniente do Kaggle, unido a dois datasets coletados pelo grupo, o primeiro destes de músicas populares mas que não fizeram sucesso no TikTok, e o segundo de músicas não populares dentro e fora do TikTok.

Usando o PCA sobre os dados normalizados, pode-se verificar novamente que cinco componentes principais eram necessárias para explicar cerca de 64% a variância dos dados. Depois disso, os dados foram transformados usando as componentes principais, e um *plot* dos resultados obtidos foi gerado, no qual diferem-se por cores cada um dos três grupos estudados. Esse resultado se encontra na figura 8.

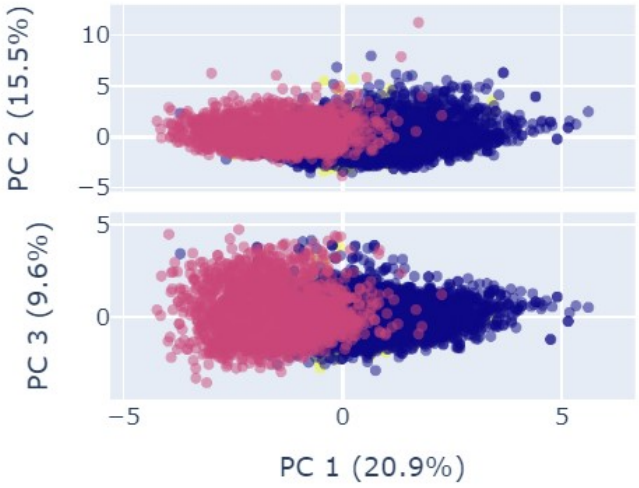

Figura 8. Gráfico de dispersão dos dados de acordo com a primeira e segunda componentes principais.

Para facilitar a visualização dos resultados, foi plotado também um gráfico somente com as músicas populares, que pode ser visualizado na figura 9.

Ele explicita que as músicas populares coletadas são bem semelhantes as músicas que fazem sucesso no TikTok, indicando haver pouca a nenhuma diferença entre as características dessas músicas.

Por fim, foram plotadas somente as músicas populares no TikTok, versus as não populares, que pode ser verificado na figura 10. Nota-se que a única diferença captu-

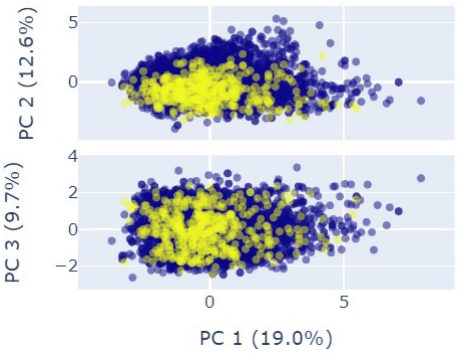

Figura 9. Gráfico de dispersão dos dados de acordo a primeira, segunda e terceira componentes principais

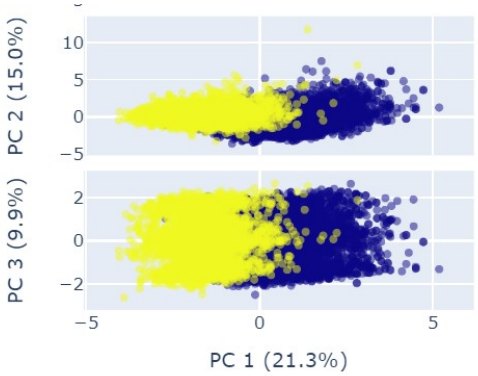

Figura 10. Gráfico de dispersão das músicas populares e não populares de acordo com a primeira, segunda e terceira componentes principais

rada pelo PCA foi a entre esses dois grupos, e ainda assim verifica-se também uma certa intercessão entre eles, indicando uma similaridade entre algumas características.

Para explicar essa diferença é necessário contextualizar a origem das músicas não populares: elas são provenientes de uma playlist de músicas de rock gótico do Spotify. Sua intercessão com as músicas populares não surpreende, afinal, existem músicas de rock que *hitaram* no TikTok.

Agora, para adentrar na diferença entre os dois grupos, é necessário verificar os pesos de cada atributo na componentes principais: a primeira componente tem como variável de maior peso a **dançabilidade** e a segunda componente tem como variável de maior peso a **energia**.

Essas variáveis definitivamente tem grande variabilidade nos grupos estudados, afinal, como citado anteriormente, muitas músicas que fazem sucesso no TikTok estão relacionadas a *trends* de danças. Ou seja, há bastantes músicas dançáveis nesse grupo. Por outro lado, as músicas góticas tem uma dançabilidade menor. Já o fator energia mede a intensidade da música. Músicas energéticas tendem a ser rápidas e barulhentas. Vemos que é principalmente em relação a segunda componente que há uma maior intercessão entre os grupos. Isso faz sentido, pois rock é um estilo de música frequentemente energético, além de que muitas das músicas que fazem sucesso no TikTok, mesmo não sendo de rock também possuem essa característica, como músicas pop.

4.5. Clusterização

O último processo de análise realizado foi uma clusterização. A ideia para a clusterização partiu da hipótese que, apesar da análise até então não ser muito conclusiva, indicando que não existia nenhuma relação muito clara entre a popularidade das músicas no TikTok e suas outras características, pode ser que, ao comparar essas músicas com músicas de outros contextos, uma relação diferente seja obtida.

Sendo assim, dois grupos diferentes de músicas foram separados e coletados para incrementar o dataset original. O primeiro grupo é composto de músicas populares que não estão presentes no dataset, ou seja, são músicas que fizeram sucesso, mas não no

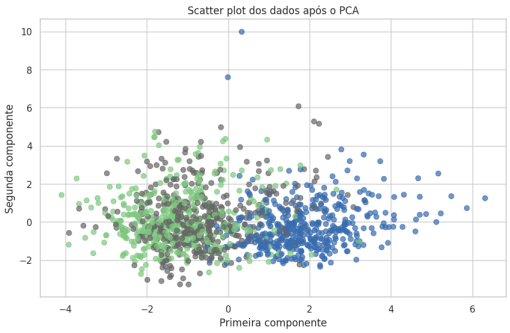

Figura 11. Gráfico de dispersão dos dados agrupados após o PCA.

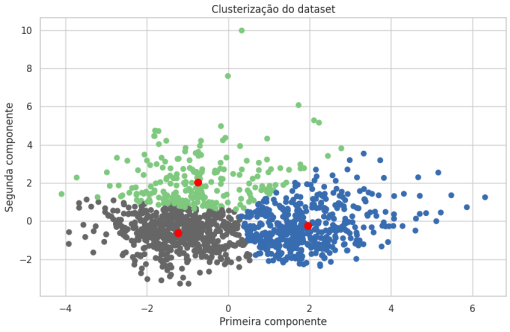

Figura 12. Resultado da clusterização sobre os dados agrupados.

TikTok. O outro grupo, por sua vez, é composto de músicas pouco populares no Spotify e que também não fazem parte do dataset (não fizeram sucesso no TikTok).

Antes da clusterização, os grupos vão ser unidos, armazenando em um atributo o grupo a que cada observação faz parte. Após isso, para obter os melhores atributos que explicam os dados, o processo de análise de componentes principais (PCA), para reduzir a dimensionalidade do dataset e obter as duas componentes que melhor explicam a disposição das observações, foi realizado e, com isso, os dados poderão ser plotados no plano e clusterizados.

Como um dos grupos de músicas era bem menor que os outros, aproximadamente 424 músicas, os outros dois grupos foram reamostrados de maneira aleatória, sendo que cada amostra possui 424 observações também. Dessa forma, o resultado obtido pela clusterização é mais justo e significativo.

Assim, as figuras 11 e 12 mostram, respectivamente, o *plot* dos dados após o PCA, onde cada cor representa a pertinência a um grupo diferente, e o mesmo gráfico de dispersão anterior mas agora com as cores alteradas para exibir as *labels* obtidas através da clusterização.

A análise de agrupamento dos dados foi feita utilizando 3 centros diferentes, onde o resultado ideal para o agrupamento seria algo próximo do gráfico de dispersão original. Esse cenário indicaria que os dados eram bem separados e, de fato, existe algo que difere cada grupo de músicas analisados.

No entanto, o resultado parcial obtido foi que há, de fato, uma interseção relativamente significativa entre grupos. Mais especificamente, existe uma interseção grande entre as observações do dataset original e as músicas populares que não fizeram sucesso no TikTok/não estão no dataset original.

Esse resultado reforça a ideia que caso exista um conjunto de características comum às músicas que fazem sucesso na rede social, esse conjunto não está sendo bem capturado pelos dados do dataset.

5. Conclusão

Nossa análise se mostrou de certa forma inconclusiva, pois não foi possível determinar um atributo específico ou mesmo um conjunto de atributos que influenciem na popularidade

de uma música na rede social escolhida. Esse é um indício de que, na verdade, os atributos observados não são tão relevantes para o sucesso da música nessa plataforma, mesmo que não seja possível afirmar isso com certeza. Pode existir alguma relação que simplesmente não foi encontrada nesse trabalho. Apesar disso, algumas hipóteses podem ser levantadas sobre as análises que fizemos.

Se existir de fato alguma relação entre os atributos observados e a popularidade da canção, ela provavelmente não é linear ou o dataset escolhido não conseguiu representar bem o universo de músicas populares na rede. Dado que este dataset tem músicas de vários gêneros diferentes, podemos questionar se este não foi um fator de confusão. Os atributos observados podem variar muito de acordo com o gênero da música. Um funk e um rock, por exemplo, podem ter atributos muito divergentes, mas a mesma popularidade na plataforma. Logo a dispersão desses atributos por causa de diferentes gêneros pode tornar a análise mais complexa. Talvez entre músicas de um gênero específico exista uma relação bem mais clara.

Outra questão é o fator humano. O TikTok é uma rede social usada por milhões de pessoas diariamente e se tornou um veículo de divulgação de artistas, ou seja, se um cantor lançar uma música e ela se popularizar no TikTok, a chance dessa música ter sucesso na vida real é muito alta. Dessa forma, existem muitos artistas que lançam músicas e criam coreografias que são relativamente fáceis de repetir, ou *trends*, e isso faz com que muitos usuários utilizem aquela música para repetir a dança ou a *trend* resultando em uma melhor divulgação dessa música. Alguns também tem contato com usuários muito famosos na rede e pedem para que estes, com maior influência, usem a música, o que também é um fator importante. Em suma, talvez a receita para o sucesso não esteja na música em si, mas na relação entre as pessoas que usam a rede.

Estas são, obviamente, apenas algumas hipóteses levantadas quando observamos que não foi encontrada nenhuma relação óbvia entre os atributos das músicas e sua popularidade. De qualquer forma, podemos concluir que são necessários mais estudos sobre esse assunto antes que se possa afirmar que existe (ou que não existe) uma receita para o sucesso.

Referências

[Carolina Walliter 2021] Carolina Walliter (2021). TikTok no Brasil e na sua marca: 10 estatísticas para arrasar em 2022. `https://www.shopify.com/br/blog/tiktok-brasil`. Online; acesso em 15 de Dezembro de 2022.

[Kaggle 2022] Kaggle (2022). Kaggle: Your Machine Learning and Data Science Community. `https://www.kaggle.com/`. Online; acesso em 02 de Dezembro de 2022.

[Spotify 2022] Spotify (2022). Spotify for Developers. `https://developer.spotify.com/discover/`. Online; acesso em 02 de Dezembro de 2022.