Clusterização para recomendação de músicas para um usuário no *Spotify* utilizando o *software* R

Lorayne Veri
Programa de Pós-Graduação em Engenharia de Produção - Universidade Federal do Paraná
Mariana Kleina
Programa de Pós-Graduação em Engenharia de Produção - Universidade Federal do Paraná
Marcell Mariano Correa Maceno
Programa de Pós-Graduação em Engenharia de Produção - Universidade Federal do Paraná
Paola Andrea Rico Belalcazar
Programa de Pós-Graduação em Engenharia de Produção - Universidade Federal do Paraná

Resumo: A forma de consumir música vem mudando ao longo dos anos e com a chegada dos aplicativos de músicas, esta mudança está mais acelerada. Por sua vez, a internet fez com que o consumo se tornasse altamente dinâmico, trazendo inúmeras vantagens para os consumidores e produtores. Atualmente, uma das maneiras de consumir música é a partir das plataformas digitais em *streaming*. Neste contexto, o presente artigo tem como objetivo examinar os conteúdos musicais dos usuários e buscar recomendar novas opções musicais de acordo com as análises obtidas por meio dos dados do *Spotify* utilizando sua API e bibliotecas do *software* R.

Palavras-chave: *Software* R, Clusterização, *Spotify*, Recomendação.

Clustering for music recommendation for a user on Spotify using R software

Abstract: The way of consuming music has been changing over the years and with the arrival of music apps, this change is more accelerated. In turn, the internet has made consumption highly dynamic, bringing numerous advantages to consumers and producers. Currently, one of the ways to consume music is through digital streaming platforms. In this context, this article aims to examine users' musical content and seek to recommend new musical options according to the analysis obtained through Spotify data using its API and R libraries.

Keywords: *Software* R, Clustering, Spotify, Recommendation.

1. Introdução

Na segunda metade do século XX, a prática da comercialização de músicas e de grupos musicais entrou em ascensão ganhando forma, até se tornar a indústria fonográfica que conhecemos hoje. A perspectiva agora dentro do mercado fonográfico é que a música, o grupo ou o artista são itens comerciais, ou seja, o sucesso do produto se baseia em sua popularidade e quanto o artista consegue reverter sua arte em fatores financeiros e midiáticos. Além disso, com o desenvolvimento tecnológico, a forma de exploração da música foi mudando, se aproximando cada vez mais do cotidiano da sociedade.

Neste cenário do mercado fonográfico, o *Spotify*, um serviço de *streaming* de música idealizado por Daniel Ek e Martin Lorentzon em 2006, surgiu com o propósito de proporcionar uma nova forma de ouvir música, além da promessa de combater a pirataria, seguindo o fluxo de uma revolução. Sendo lançado em 2008, para diversos dispositivos, a plataforma fornece conteúdo por meio da transmissão instantânea de dados de áudio e vídeo por meio de redes, serviço conhecido como *streaming*.

No ano de 2014 o *Spotify* passou a incluir algoritmos que possibilitam recomendações de conteúdo personalizado com o objetivo de aprimorar constantemente a qualidade do serviço (FRANZONI, 2017).

Desta forma, o *Spotify,* um dos maiores acervos de música *online*, promete a possibilidade de ouvir e descobrir músicas gratuitamente (mesmo com uma versão paga mensalmente) e a disponibilidade de um serviço de *playlists* e rádios personalizadas.

Sendo assim, o enfoque deste artigo é um algoritmo de recomendação com abordagem baseada em clusterização usando o método de agrupamento por meio da amostra do banco de dados. Será abordado desde a extração desses dados pela API (interface de programação de aplicações) do *Spotify*, passando para a manipulação dos dados e realizando a clusterização por meio da linguagem de programação R e, finalmente, serão discutidos os resultados.

2. Revisão da Literatura

Nesta seção são apresentadas as ferramentas necessárias para o entendimento e execução da pesquisa.

2.1. Plataforma de *streaming Spotify*

O aplicativo de músicas *Spotify* é um serviço de *streaming* digital que foi desenvolvido em 2006 e lançado no dia 7 de outubro de 2008 por meio da *startup* sueca *Spotify AB.* Ele está disponível em 65 regiões do mundo e com uma base de dados, em 2018, de aproximadamente 35 milhões de músicas (MATIAS, 2018). No Brasil, o lançamento ocorreu no mês de maio de 2014, 6 anos após o lançamento oficial do aplicativo (ARAÚJO; OLIVEIRA, 2014).
O aplicativo *Spotify*, tem como objetivo oferecer um catálogo de músicas aos seus usuários sem a necessidade de adquirir algo físico como discos, CDs ou ter que baixar os arquivos (RIOS, 2015).
No mercado global de assinatura de *streaming* de música o *Spotify* aparece na liderança, com 31% dos assinantes, e a *Apple Music* é o segundo colocado no ranking de preferência dos usuários, com cerca e 15% do mercado (LISBOA, 2022). Portanto, o *Spotify* detém, com folga, o posto de maior empresa de *streaming* musical do mundo, apesar do crescimento dos concorrentes.

2.1.1. API *Spotify*

O termo API - *Aplication Programming Interface*, cuja tradução para o português significa interface de programação de aplicações", é um conjunto de padrões e rotinas de programação para o acesso a uma plataforma *web* ou aplicativo de *software* (HUGHES, 2015). Na aplicação deste artigo, além do histórico das músicas escutadas pelo usuário no período de um ano, os dados de análise serão extraídos da API que o *Spotify* oferece gratuitamente.
A partir API, os próprios usuários são capazes de desenvolver aplicativos que oferecem novos serviços relacionados à representação musical, incluindo acesso a todos os pontos

finais e suporte para autorização do usuário. Para obter detalhes sobre os recursos, recomenda-se a leitura da documentação da API da *Web* do *Spotify* (LAMERE, 2014).

2.2. Clusterização

A clusterização é um método com o objetivo de distribuir os dados em clusters onde cada item aparece em apenas em um dos grupos. Para realizar a clusterização, existem diversas técnicas e cada uma possui suas vantagens e desvantagens (UNGAR; FOSTER, 2000).

2.2.1. Clusterização *k-means*

Os algoritmos não supervisionados são extremamente competentes em clusterização, ou agrupamento de dados, possibilitando inferências sobre uma grande quantidade de dados sem ser necessária interação humana. Sendo assim, o *k-means* é um método de clusterização muito popular em análise e dados e *data mining*, principalmente por sua precisão e facilidade de implementação (HORNIK, 2017).
Na clusterização do *k-means*, existem algumas limitações conhecidas, pode-se citar: conhecimento a priori dos dados, para a escolha do número *k* de grupos que será gerado; os grupos gerados no final são muito sensíveis a escolha inicial dos centroides; pode produzir grupos vazios (RICCI, 2010).
O método é composto por quatro etapas. Primeiramente é feita a escolha aleatória de *k* centroides dos pontos de amostra como centros iniciais do cluster. Em seguida, a atribuição de cada amostra ao centroide mais próximo. Na terceira etapa, a alteração dos centroides para o centro das amostras que lhe foram conferidas. Por fim, a repetição da segunda e terceira etapa, até que a atribuição do cluster não seja alterada ou uma tolerância seja definida pelo usuário ou um número máximo de iterações seja realizado (RASCHKA, 2015).
O resultado é um agrupamento cuja performance depende diretamente da qualidade dos dados e do *k* definido. Nota-se que o processo de escolha do *k* pode não ser trivial.

2.3. Pareto

O princípio de Pareto, nomeado em homenagem ao economista Vilfredo Pareto, é mais conhecido nos círculos da qualidade como a regra 80/20. É uma ferramenta importante para a identificação dos problemas, pois elabora gráficos que permitem uma melhor visualização dos dados (BEZZERA, 2019).
Desta forma, a ferramenta aponta os pontos críticos na organização, tais como, erro na produção, na qualidade ou desperdício de materiais. Portanto, o objetivo é criar um gráfico que auxiliará, mostrando de forma decrescente, os processos que causam maior efeito para a empresa, esse método auxilia na tomada de decisões (FALCONI, 2009).
De acordo com Falconi (2009), a análise de Pareto é uma ferramenta muito simples e poderosa para o gerente, pois facilita e ajuda a classificar e priorizar os seus problemas. Por exemplo, se o gerente optar por reduzir o nível de estoque da empresa, ele pode utilizar uma análise de Pareto, que demonstrará que poucos itens são responsáveis pela maior parte do capital estocado. O princípio de Pareto é um método universal para separar os problemas em duas classes: os poucos vitais e muitos triviais.

3. Metodologia

A base de dados escolhida para o desenvolvimento do trabalho, foi o histórico das músicas ouvidas no período de 22 de agosto de 2021 a 22 de agosto de 2022 de um usuário do *Spotify* e uma base construída a partir de dados coletados na *Web* API, ambos da plataforma *Spotify*. Para a manipulação e análise dos dados obtidos foi utilizado o *software* R e o *k-means* para seu agrupamento. Na sequência é explicada cada uma das etapas do trabalho.

3.1. Conjunto de dados utilizados

Nesta seção serão descritas as bases de dados utilizadas para o estudo.

3.1.1. Base de dados do perfil

Ao ser solicitado, a plataforma *Spotify* disponibiliza ao usuário o histórico das músicas escutadas no período de um ano da data solicitada pelo perfil, assim para este trabalho, foi solicitado o histórico de um usuário que disponibilizou o uso de seus dados, o qual foi a principal base de dados utilizada. O arquivo continha 10.011 observações, com 4 variáveis - nome do artista da música, música, dia e hora que a música foi escutada, tempo que a música foi ouvida, porém para a análise deste trabalho, não foi necessária esta granularidade, assim, os dados foram agrupados em nome do artista da música, música e tempo que a música foi ouvida, gerando 2.442 observações. A Tabela 1 apresenta um descritivo das variáveis usadas no trabalho.

Tabela 1 – Variáveis da base de dados do perfil

Nome da variável	Significado	Tipo
artistName	Nome do artista de cada música	*String*
trackName	Nome das músicas	*String*
msPlayed	Quantidade de milésimos de segundos que uma faixa foi reproduzida	Número real maior ou igual a zero

Fonte: Autoria própria com base nos dados de Spotify (2023)

Para obter a amostra de estudo do histórico de *streaming*, proceder-se-á fazer uma análise de Pareto (80%) com a finalidade de ter uma amostra final que contenha as músicas mais representativas (maiores valores da variável *msPlayed*), ou seja, as mais ouvidas pelo perfil selecionado, gerando uma amostra com 620 músicas.

3.1.2. Base de dados para a recomendação

Em concordância com o objetivo do trabalho, deve-se obter a base de dados das músicas que poderão ser recomendadas para o usuário. Esta base foi construída a partir das músicas mais tocadas no Brasil, com 70% dos artistas mais escutados no histórico de dados do usuário (seção 3.1.1). A função utilizada para esta coleta foi *getTopTracks* do pacote *RSpotify*, do *software* R, por meio da API do *Spotify*. A base final das músicas para a recomendação contém 8 variáveis e 726 observações, das quais, 4 variáveis foram as mais relevantes para o estudo e são apresentadas na Tabela 2.

Tabela 2 – Variáveis da base de dados para a recomendação

Nome da variável	Significado	Tipo
track_id	Identificação das músicas no *Spotify*	*String*
name	Nome dos itens reproduzidos (músicas)	*String*
artist_name	Nome dos artistas	*String*
artist_id	Identificação dos artistas no *Spotify*	*String*

Fonte: Autoria própria com base nos dados de Spotify (2023)

3.2. Variáveis da clusterização

Para poder classificar as músicas, foram escolhidas as variáveis da clusterização com base nas características próprias das músicas. Por isso, foi usada a função *getFeatures* do pacote *RSpotify*, do *software* R, que obtém a informação da base de dados da *Web* API. Gerou-se 13 variáveis para os dois conjuntos de dados descritos anteriormente (seção 3.1.1

e seção 3.1.2) que contém propriedades com as quais é possível encontrar semelhanças entre elas, das quais somente foram selecionadas 6 que são as mais relevantes para o estudo, seu detalhamento se encontra na Tabela 3.

Tabela 3 – Parâmetros das músicas

Nome da variável	Significado	Tipo
danceability	Quanto a música é dançável	Números real entre 0 e 1, sendo 0 o valor menos dançável e 1 o mais dançável.
energy	Nível de atividade da música	Números real entre 0 e 1, sendo 0 o valor com menor energia e 1 com maior energia.
tempo	Compasso da música	Número real
valence	Detecta a positividade, sentimento de felicidade da música	Números real entre 0 e 1, sendo 0 o valor com menor positividade e 1 o valor com maior positividade.
acousticness	Indica o quanto a música é acústica	Números real entre 0 e 1, sendo 0 o valor com menor acústica e 1 o com maior acústica.
liveness	Indica a presença de plateia	Números real entre 0 e 1, sendo 0 o valor com menor plateia e 1 o com maior plateia.

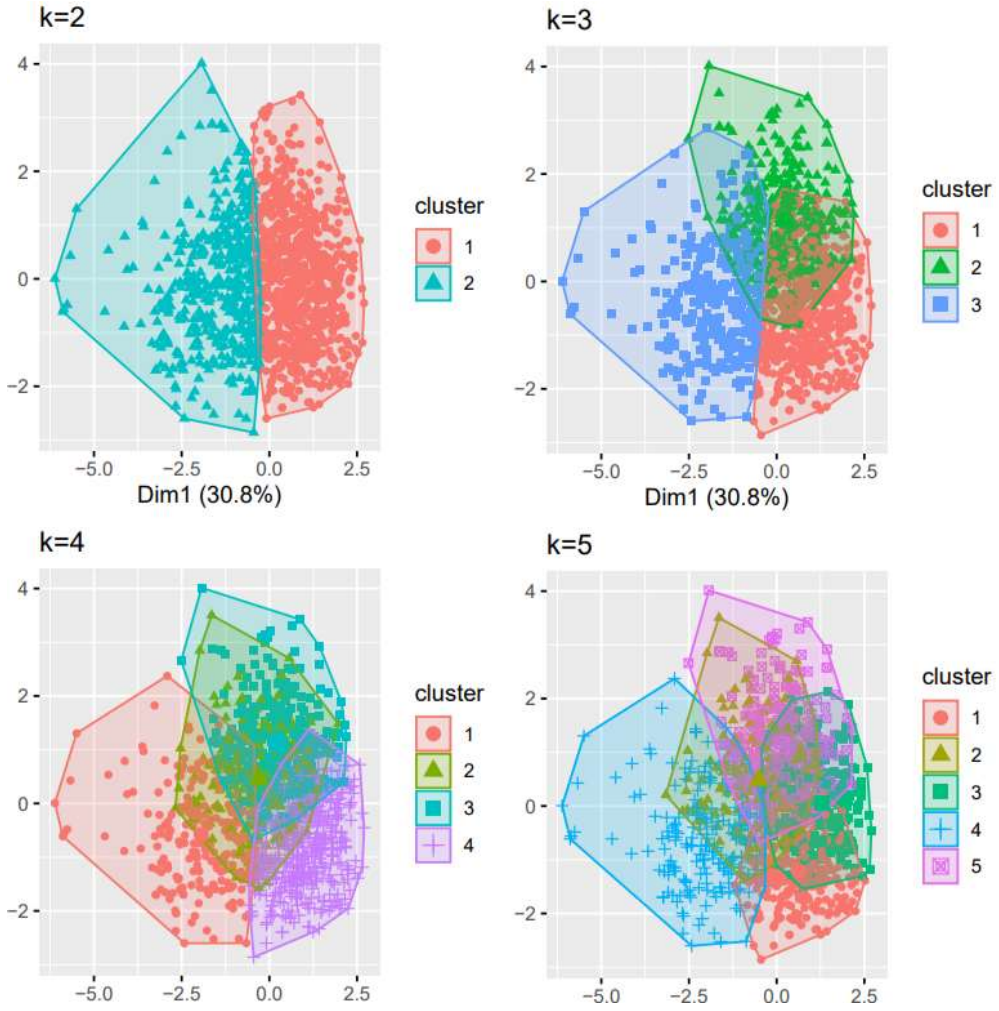
Fonte: Autoria própria com base nos dados de Spotify (2023)

4. Resultados

Na amostra escolhida pelos autores deste artigo, a música mais escutada pelo usuário selecionado foi *Don't Stop Believin – Journey*, 2,54 horas e a menos escutada foi Seu Astral – Jorge e Mateus, 6,53 minutos.

Após a coleta dos dados e a definição dos parâmetros, foi realizada a padronização dos dados, via método de *scale*, que consiste em subtrair os valores de sua média e dividir pelo desvio padrão, e posteriormente executado o método de clusterização *k-means*. Este método tem uma característica importante, nele é necessário definir o número de clusters. Na Figura 1 pode-se observar como ficam distribuídos os clusters em quatro cenários diferentes, quando são considerados 2, 3, 4 e 5 centroides com 100 iterações.

Figura 1 – Clusterização com diferentes números de centroides

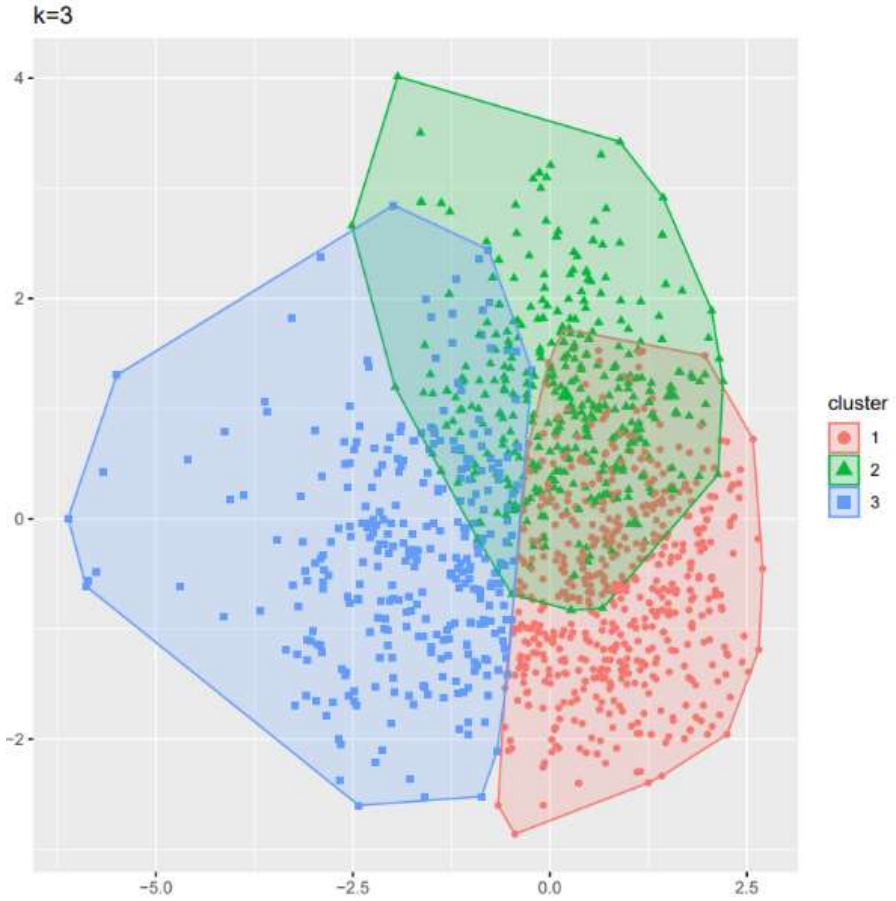


Fonte: Autoria própria (2023)

Definir o número de clusters a serem considerados pode ser uma tarefa bastante difícil. A análise deste trabalho terá será realizada com *k*=3, escolhido visualmente pelos autores da pesquisa.

Na clusterização realizada utilizando 3 centroides, obteve-se 587 músicas no cluster 1, em vermelho, 420 músicas no cluster 2, em verde, e 339 músicas no cluster 3, em azul (Figura 2). Os pontos localizados na intersecção dos clusters são músicas que possuem características, variáveis, em comum.

Figura 2 – Clusterização com 3 centroides



Fonte: Autoria própria (2023)

Na Tabela 4 pode-se ver a distribuição das músicas do histórico de dados do usuário e as músicas a serem recomendadas em cada cluster.

Tabela 4 – Resultado da clusterização com 3 centroides

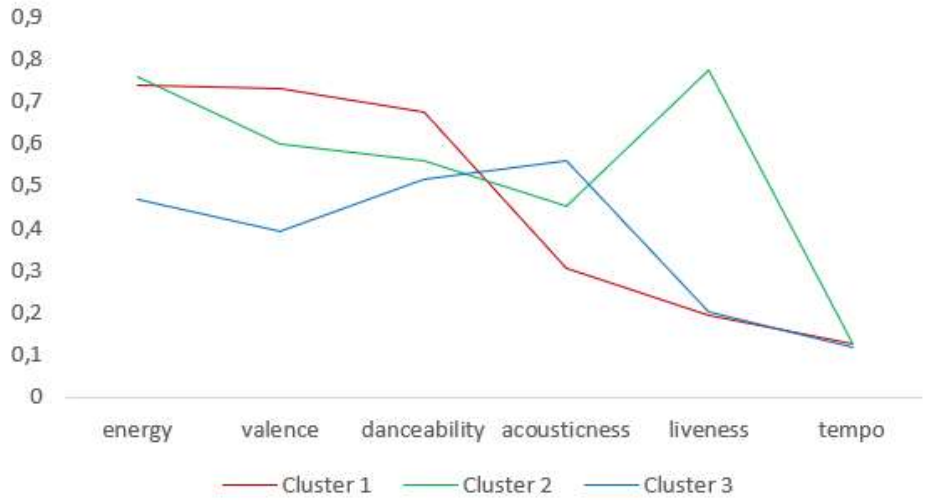
Item	Cluster 1	Cluster 2	Cluster 3	Total
Músicas do histórico do usuário	284	188	148	620
Músicas a serem recomendadas	303	232	191	726
Total de músicas	587	420	339	1.346

Fonte: Autoria própria com base nos dados do *Spotify* (2023)

Observou-se que entre todas as músicas agrupadas no cluster 1, 48,3% delas são músicas contidas no histórico do último ano do usuário, já no cluster 2, estas músicas representam 44,7% e no cluster 3 são 43,6%. Assim, entende-se que as músicas agrupadas no cluster 1 irão agradar mais o usuário, visto que neste grupo existem mais músicas que o usuário escutou recentemente.

Com a Figura 3, pode-se analisar a média das variáveis escolhidas em cada cluster. O cluster 1, recomendado para o usuário, possuiu a maior média nas variáveis *energy*, que informa o nível de atividade de música, e *valence*, variável que indica a positividade da música e, em contrapartida, possui menor média das variáveis de tempo, que informa o compasso da música, e *liveness*, variável que indica a presença de plateia.

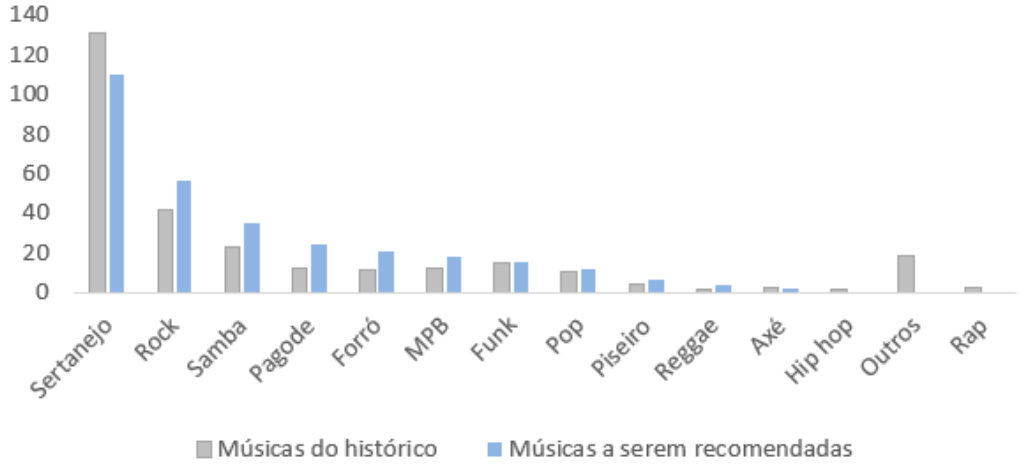
Figura 3 – Média das variáveis por cluster



Fonte: Autoria própria (2023)

Fixado o cluster 1 como recomendado ao usuário, observou-se que o estilo mais recomentado para o usuário é o sertanejo, com 36% das músicas, em seguida o rock com 18% das músicas recomendadas, conforme Figura 4.

Figura 4 – Músicas do histórico *versus* músicas recomendadas



Fonte: Autoria própria (2023)

Os artistas mais recomendados para o usuário foram Bezerra da Silva e *Mc* Kevinho, ambos com 9 músicas cada, e em seguida *Red Hot Chili Peppers* com 8 músicas recomendadas.

4. Conclusões

No momento atual, as quantidades massivas de dados que geramos a partir das plataformas de *streaming* e em pesquisas na internet traz a necessidade de selecionar músicas relevantes de acordo com o perfil do usuário, assim o objetivo deste artigo foi propor um algoritmo de recomendação a partir das músicas já escutas pelo usuário na plataforma de *streaming* do *Spotify*.

Este artigo apresentou um sistema de recomendação com abordagem baseada em clusterização usando o método de agrupamento, por meio do histórico de músicas ouvidas pelo usuário no período de um ano no *Spotify* e da amostra do banco de dados do *streaming*

coletada com a ajuda da função *getTopTracks* do pacote *RSpotify*, do *software* R. A amostra do artigo foi filtrada a partir da análise de Pareto com a finalidade de ter uma amostra final que contenha as músicas mais representativas. Após a coleta dos dados e a definição dos parâmetros, foi realizada a padronização dos dados e executado o método de clusterização *k-means.*

Para a análise do algoritmo posposto, o perfil escolhido possui como característica ser eclético. Além desta afirmação ser confirmada analisando os dados do histórico das músicas escutadas no período de um ano pelo usuário, obtidos do *Spotify*, o resultado da clusterização refletiu esta característica, pois não se tem uma grande diferença do percentual de músicas já ouvidas pelo usuário nos clusters 1, 2 e 3, com 48,3%, 44,7% e 43,6%, respectivamente. Como o cluster 1 obteve o maior percentual de músicas já ouvidas pelo usuário, as músicas com maiores valores de *energy* e *valence*, agrupadas neste cluster, foram as recomendadas para o perfil selecionado.

Por fim, o estilo de música mais recomendado para o usuário, de acordo com os dados coletados, é o sertanejo e os artistas mais recomendados foram Bezerra da Silva e *Mc* Kevinho, ambos com 9 músicas cada, e em seguida *Red Hot Chili Peppers* com 8 músicas recomendadas.

Para o desenvolvimento desta pesquisa foi utilizada apenas a amostra de um usuário do aplicativo, porém a metodologia desenvolvida pode ser aplicada para qualquer amostra.

Referências

ARAÚJO, L.; OLIVEIRA, C. **Música em fluxo: experiências de consumo musical em serviços de streaming.** Temática, [s.l.], v. 10, n. 10, p. 122-137, 2014. Mensal. Disponível em <http://www.periodicos.ufpb.br> Acesso em: 22 ago. 2022.

BEZERRA, F. **Diagrama de Pareto: O que é e como fazer.** Disponível em:<http://www.portal-administracao.com/2014/04/diagrama-de-pareto-passo-apasso.html> Acesso em: 22 ago. 2022.

FALCONI, V. **TQC: Controle da qualidade total no estilo Japonês.** Nova Lima/MG: Falconi, 2009.

FRANZONI, T. **Motivação de consumo de música por streaming dos usuários do Spotify.** 2017.

HORNIK, K. 2017. Disponível em <**https://cran.rproject.org/doc/FAQ/R-FAQhtml**> Acesso em 20 ago. 2022.

HUGHES, C. **Understanding the Spotify Web API. Spotify Labs**, 9 mar. 2015. Disponível em: <https://labs.spotify.com> Acesso em: 20 ago. 2022.

LAMERE, P. spotipy. 2014. Disponível em: <**https://spotipy.readthedocs.io**>. Disponível em: Acesso em 20 ago. 2022.

LISBOA, A., 2022. Editado por Douglas Ciriaco. **Os apps de música por streaming mais usados no mundo.** Disponível em <https://canaltech.com.br/apps/os-apps-de-musica-por-streaming-mais-usados-no-mundo-207147/>. Disponível em: Acesso em 22 ago. 2022.

MATIAS, A. **Spotify, 10 anos: como o serviço de streaming mudou a música.** Disponível em <https://reverb.com.br/artigo/spotify-10-anos-como-o-servico-destreaming-mudou-a-musica>. Acesso em: 20 ago. 2022.

RASCHKA, S. **Python Machine Learning: Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics.** Birmingham: Packt Publishing Ltd, 2015.

RICCI, F., ROKACH, L., SHAPIRA, B., and Kantor, P. B. **Recommender Systems Handbook.** Springer-Verlag New York, Inc., New York, NY, USA, 1st edition. 2010.

RIOS, R. **Spotify: Streaming e as novas formas de consumo na era digital.** In: CONGRESSO DE CIÊNCIAS DA COM. NA REGIÃO NORDESTE, 17., 2015.Anais eletrônicos. Disponível em <http://www.portalintercom.org.br> Acesso em: 22 ago. 2022.

UNGAR, L., P. FOSTER, D. **Clustering methods for collaborative filtering.** 2000.