**Julia Hellen Franco Ferreira**

# Sistema de Recomendação baseado em Filtragem Colaborativa utilizando dados binários do Spotify

Niterói - RJ, Brasil

19 de dezembro de 2023

Universidade Federal Fluminense

**Julia Hellen Franco Ferreira**

# Sistema de Recomendação baseado em Filtragem Colaborativa utilizando dados binários do Spotify

**Trabalho de Conclusão de Curso**

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Prof. Dr. Jony Arrais Pinto Junior

Co-Orientador(a): Profa. Dra. Jessica Quintanilha Kubrusly

Niterói - RJ, Brasil

19 de dezembro de 2023

Universidade Federal Fluminense

**Julia Hellen Franco Ferreira**

# Sistema de Recomendação baseado em Filtragem Colaborativa utilizando dados binários do Spotify

Monografia de Projeto Final de Graduação sob o título *"Sistema de Recomendação baseado em Filtragem Colaborativa utilizando dados binários do Spotify"*, defendida por Julia Hellen Franco Ferreira e aprovada em 19 de dezembro de 2023, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

---

**Prof. Dr. Jony Arrais Pinto Junior**
Departamento de Estatística – UFF

---

**Profa. Dra. Jessica Quintanilha Kubrusly**
Departamento de Estatística – UFF

---

**Profa. Dra. Karina Yuriko Yaginuma**
Departamento de Estatística – UFF

---

**Prof. Dr. Rafael Santos Erbisti**
Departamento de Estatística – UFF

Niterói, 19 de dezembro de 2023

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

F825s Franco Ferreira, Julia Hellen
Sistema de Recomendação baseado em Filtragem Colaborativa utilizando dados binários do Spotify / Julia Hellen Franco Ferreira. - 2023.
49 f.: il.

Orientador: Jony Arrais Pinto Junior.
Coorientador: Jessica Quintanilha Kubrusly.
Trabalho de Conclusão de Curso (graduação)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2023.

1. Sistema de Recomendação. 2. Dados Binários. 3. Filtragem Colaborativa. 4. Produção intelectual. I. Arrais Pinto Junior, Jony, orientador. II. Quintanilha Kubrusly, Jessica, coorientadora. III. Universidade Federal Fluminense. Instituto de Matemática e Estatística. IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

# Resumo

A industria musical é responsável por uma imensidão de cantores, gêneros e ritmos que produzem uma infinidade de músicas, dificultando muitas vezes ao processo de selecionar, filtrar e organizar estes dados conforme as preferências do consumidor. Os Sistemas de Recomendação buscam sugerir itens baseados na semelhança de outros itens ou baseados nas preferências de outros consumidores com gostos semelhantes para resolver este probelma. Dado o grande volume de dados encontrado nas plataformas de *streamings* de músicas como o *Spotify*, estes Sistemas de Recomendação são uma ferramenta essencial para garantir de forma automatizada e personalizada a indicação de músicas para melhorar a experiência do cliente e garantir uma maior satisfação. Este estudo desenvolveu um Sistema de Recomendação por Filtragem Colaborativa em uma base de dados binários que representa o consumo de músicas em *playlists* do Spotify. Diversas metodologias foram exploradas, destacando a Similaridade entre Usuários (*playlists)* por meio de medidas como *Jaccard*, *3W-Jaccard*, *Sorensen-Dice*, *Ochai*, *Sokal & Sneath*, e *Sokal & Michener*, junto com diferentes valores de $k$ para identificar as *playlists* similares e em seguida recomendar músicas a elas. O $k$ é definido como o número de *playlists* similares que será utilizado ao longo dos cálculos de similaridades. A base de dados contém 869 *playlists* distintas e 34443 músicas únicas. A comparação das medidas de similaridade, feita através do método *Leave-one-out*, envolveu um cenário de simulação, destacando a eficácia da Similaridade de *3W-Jaccard* com $k = 5$. Após a definição da medida e vizinhos mais próximos, foram realizadas recomendações para três playlists, evidenciando um desempenho promissor ao sugerir músicas com gêneros semelhantes ou idênticos. A conclusão destaca a significância de levar em conta as peculiaridades associadas às entradas de valor zero, que podem ser interpretadas como a irrelevância de uma música para a *playlist* ou simplesmente o desconhecimento da mesma. Além disso, ressalta-se a importância de uma escolha criteriosa da medida de similaridade e do número $k$ de vizinhos para garantir resultados satisfatórios.

Palavras-chave: Filtragem Colaborativa. Sistemas de Recomendação. Dados Binários. Leave-one-out. Spotify.

# Dedicatória

Dedico este trabalho ao meu irmão Jorge Lucas,

Ainda que seja mais novo que eu, não consigo lembrar da minha vida sem a sua presença. Minha jornada, desde a infância até as alegrias atuais, é indissociável da sua existência.

Você é mais do que um irmão, é uma fonte inesgotável de inspiração, perseverança e dedicação. Suas qualidades moldaram não apenas a nossa relação, mas também influenciaram diretamente a pessoa que me tornei.

Obrigado por ser meu companheiro de vida, por compartilhar risadas e desafios, e por ser a luz constante nos momentos de escuridão. Este trabalho é dedicado a você, meu irmão, em reconhecimento e gratidão pela sua inestimável contribuição para a minha jornada.

# Agradecimentos

Gostaria de expressar minha profunda gratidão a todas as pessoas e instituições que desempenharam um papel fundamental na realização deste Trabalho de Conclusão de Curso. Este projeto alcançou o sucesso devido ao apoio, orientação e estímulo de diversos indivíduos e entidades.

Em primeiro lugar, quero agradecer aos meus dedicados orientadores, Jony Arrais Pinto Júnior e Jéssica Quintanilha Kubrusly. Sua orientação, paciência e sugestões valiosas foram essenciais para o desenvolvimento e conclusão deste trabalho.

Expresso meu reconhecimento aos professores Karina Yuriko Yaginuma e Rafael Santos Erbisti, membros da banca examinadora. Suas valiosas contribuições e a avaliação criteriosa enriqueceram substancialmente este estudo, conferindo-lhe qualidade e relevância elevadas.

À minha querida família - minha mãe, Ilca Flávia Franco, a mulher que me ensinou a nunca desistir dos meus sonhos; meu pai, Jorge Wilson Ferreira, que me ensinou a usar a comunicação oral ao meu favor; e meu amado irmão, Jorge Lucas Ferreira, que me ensinou a ver a beleza da vida mesmo no meio do caos -, expresso meu sincero agradecimento pelo apoio incondicional. O incentivo emocional e o suporte prático foram fundamentais para superar os desafios inerentes a este percurso acadêmico.

Aos colegas de curso, Gabriela Barros, Filipe Nascimento e Thiago Lima, agradeço pela colaboração, compartilhamento de conhecimentos e amizade ao longo dessa jornada. Ao final deste processo, percebo que não apenas foram colegas, mas também verdadeiros amigos.

Aos meus colegas da Globo, que passaram conhecimentos e discussões enriquecedores necessários para a realização deste trabalho, agradeço pelo ambiente propício à pesquisa e aprendizado.

Quero dedicar este momento especial à memória da tia Valeria de Barros, uma mulher amorosa, atenciosa e amiga. Tenho certeza de que, se estivesse presente, ela celebraria esta conquista ao meu lado, junto com sua filha, Gabriela Barros. Da mesma forma,

dedico este feito ao meu avô David Amador, que teria imenso orgulho de cada passo da minha jornada.

Uma menção especial vai para minha maior fonte de inspiração, Beyoncé Knowles-Carter. Suas canções não apenas me acompanharam nos momentos difíceis, mas também celebraram comigo grandes vitórias. Além disso, ela desempenhou um papel crucial em me ajudar a compreender meu papel como mulher negra na sociedade.

Finalmente, expresso minha gratidão a todos que, de alguma forma, contribuíram para a concretização deste trabalho, inclusive meus fiéis companheiros de quatro patas, Biscoitinho e Bento, que estiveram ao meu lado nas noites de escrita. Cada forma de apoio foi fundamental, e estou sinceramente agradecido por cada contribuição.

# Sumário

**Lista de Figuras**

**Lista de Tabelas**

**1 Introdução** p. 11

**2 Materiais e Métodos** p. 16

2.1 Spotify . . . . . p. 16

2.2 Materiais . . . . . p. 17

2.3 Sistemas de Recomendação . . . . . p. 19

2.3.1 Recomendação Baseada em Conteúdo . . . . . p. 21

2.3.2 Recomendação Baseada em Filtros Colaborativos ou Cosumo . . p. 22

2.3.3 Recomendação Híbrida . . . . . p. 23

2.4 Filtragem Colaborativa . . . . . p. 24

2.4.1 Avaliação implícita e explicita . . . . . p. 25

2.4.2 Dados Binários . . . . . p. 25

2.4.3 Matriz de Utilidade . . . . . p. 27

2.4.4 Matrizes de Similaridade . . . . . p. 28

2.4.4.1 Recomendação a partir da Similaridade entre Usuários p. 29

2.4.4.2 Recomendação a partir da Similaridade entre Itens . . p. 29

2.4.5 Medidas de Similaridades . . . . . p. 30

2.4.5.1 Similaridade de Jaccard e 3W-Jaccard . . . . . p. 30

2.4.5.2 Similaridade de Sorensen-Dice . . . . . p. 32

2.4.5.3 Similaridade de Ochai . . . . . p. 33
2.4.5.4 Similaridade de Sokal & Sneath . . . . . p. 33
2.4.5.5 Similaridade de Sokal & Michener . . . . . p. 34
2.4.6 Recomendação . . . . . p. 35
2.4.6.1 *Leave-One-Out* . . . . . p. 35

**3 Análise dos Resultados** p. 38
3.1 Análise Descritiva . . . . . p. 38
3.2 Filtragem Colaborativa . . . . . p. 39
3.2.1 Comparação entre as Medidas de Similaridades . . . . . p. 39
3.2.2 Recomendação a partir da Similaridade de *3W-Jaccard* entre Usuários p. 43

**4 Conclusão** p. 46

**Referências** p. 48

# Lista de Figuras


1 Trilho "Beyoncé tem um álbum novo. Surtando?" na *Home* do consumidor p. 17

2 Trilho "Acordou cedo né?" na *Home* do consumidor . . . . . p. 17

3 Processo de Recomendação . . . . . p. 20

4 Recomendação Baseada em Conteúdo por Usuários . . . . . p. 21

5 Recomendação Baseada em Conteúdo por Item . . . . . p. 22

6 Recomendação Baseada em Filtros Colaborativos por Usuário . . . . . p. 23

7 Recomendação Baseada em Filtros Colaborativos por Item . . . . . p. 23

8 Fluxograma do Sistema de Recomendação . . . . . p. 35

9 Fluxograma da técnica Leave-one-out . . . . . p. 37

10 Estimativas para Similaridades de *Jaccard*, *3W-Jaccard*, *Sorensen-Dice*, *Ochai*, *Sokal & Sneath* e *Sokal & Michener* . . . . . p. 41

# Lista de Tabelas


1 Descrição das variáveis referente as *playlists* do *Spotify* . . . . . p. 19

2 Informações da variável *track* . . . . . p. 19

3 Matriz de Utilidade . . . . . p. 28

4 Matriz de Utilidade . . . . . p. 30

5 Estatísticas sobre as *playlists* . . . . . p. 38

6 Materiz de Utilidade: Playlists X Músicas . . . . . p. 39

7 As 5 primeiras linhas e colunas da Materiz de Similaridade . . . . . p. 43

8 As 5 *playlists* mais semelhantes em relação a *playlist Rap* . . . . . p. 43

9 As 5 músicas recomendas para *playlist Rap* . . . . . p. 44

10 As 5 *playlists* mais semelhantes em relação a *playlist Christmas Favorites* p. 44

11 As 5 músicas recomendas para *playlist Christmas Favorite* . . . . . p. 44

12 As 5 *playlists* mais semelhantes em relação a *playlist Classic* . . . . . p. 45

13 As 5 músicas recomendas para *playlist Classic* . . . . . p. 45

# 1 Introdução

A música é uma expressão artística amplamente difundida em diferentes culturas ao redor do mundo. Desde tempos remotos, as composições musicais desempenharam um papel significativo ao relatar eventos históricos e transmitir ideias relacionadas à sociedade em que vivem (NAPOLITANO, 2015). Com uma vasta diversidade de cantores, gêneros e ritmos, são criadas inúmeras músicas o que muitas vezes torna desafiador para os consumidores desta arte escolherem entre elas. Selecionar, filtrar e organizar essas informações de acordo com as preferências do público pode se mostrar uma tarefa complexa (ROZENDO, 2017).

Diante desse cenário, os Sistemas de Recomendação (SR) surgem como ferramentas que visam facilitar a escolha e personalizar a experiência dos usuários em plataformas de *streaming* de música. Baseados no princípio de sugerir itens com base na semelhança com outros ou nas preferências de consumidores que possuam gostos similares, os SRs têm como objetivo automatizar e aprimorar a indicação de itens ou conteúdos, buscando garantir uma maior satisfação dos usuários (GÓIS, 2015).

Para SILVA (2019), no contexto do Aprendizado de Máquina (*Machine Learning*), os SRs podem ser categorizados em três principais abordagens. A primeira, a **Filtragem de Conteúdo**, concentra-se no conteúdo do item ou no perfil do consumidor. Por exemplo, no domínio musical, uma música pode ser recomendada baseando-se em suas características intrínsecas como gênero, ritmo, artista e duração. Ainda na Filtragem de Conteúdo é possível usar o perfil do consumidor, como por exemplo a faixa etária, sexo, nacionalidade e localização, para realizar recomendações mais personalizadas. Em contraste, a **Filtragem Colaborativa** que foca nas interações dos consumidores, ou seja, o consumo de usuários semelhantes referente a item ou nos padrões de consumo associados a itens específicos. Nesta abordagem uma música é recomendada com base na similaridade entre padrões de consumo. Se dois ouvintes frequentemente escolhem as mesmas músicas, uma faixa apreciada por um pode ser sugerida ao outro. Além do consumo ser uma das formas de se realizar recomendações, na Filtragem Colaborativa também pode-se usar *feedbacks*,

como classificações ou frequência de reprodução. A **Recomendação Híbrida**, como o nome sugere, combina características da Filtragem de Conteúdo e Filtragem Colaborativa. Esta combinação busca superar as limitações individuais de cada abordagem e oferecer recomendações mais precisas e pertinentes. Para o escopo deste trabalho, a ênfase será colocada exclusivamente na Filtragem Colaborativa.

O *Spotify*, líder no mercado global das plataformas de streaming, com 31% de participação de usuários [1], já contava com mais de 356 milhões de usuários ativos mensais até o ano de 2021, demonstrando sua posição dominante no mercado de áudio [2]. A Filtragem Colaborativa torna-se um aliado valioso para recomendar músicas, artistas, *playlists* e *podcasts* personalizando as ofertas para os consumidores. A recomendação está presente em várias áreas da plataforma, como por exemplo na sugestão de músicas logo após o término da reprodução de uma *playlist* e na recomendação de *playlists* na página inicial com foco nas preferências do usuário.

Recomendações musicais personalizadas, como as do *Spotify*, detêm significativa relevância para os usuários, permitindo a descoberta de novas faixas, artistas e *playlists* que se alinham às suas preferências. Essa abordagem não só amplia a variedade de experiências musicais, proporcionando sensação de exclusividade, mas também fortalece o vínculo entre a plataforma e seus usuários. Ao entregar sugestões de alta relevância, a probabilidade de manutenção e conversão desses usuários em assinantes pagos aumenta. Ademais, novos artistas se beneficiam, ganhando visibilidade através das recomendações inteligentes.

A capacidade de explorar e promover nichos musicais diversifica e democratiza o cenário, favorecendo artistas menos reconhecidos e consumidores em busca de novidades. Em meio à vasta biblioteca do *Spotify*, a recomendação auxilia na otimização do tempo, facilitando a localização de músicas desejadas. As sugestões precisas e personalizadas são cruciais para o sucesso de plataformas de *streaming*, já que aprimoram a experiência do usuário, intensificam a fidelização e incrementam a descoberta de conteúdo musical diversificado.

Neste trabalho, a técnica de Filtragem Colaborativa é empregada como método principal para recomendar músicas a *playlists* com base nos dados fornecidos pelo Spotify. Estes dados são representados de maneira binária, indicando unicamente a presença ou ausência de uma dada música em uma *playlist* alvo. Esta abordagem binária, em contraste com os

[1] https://www.tecmundo.com.br/mercado/232397-spotify-segue-lider-isolado-mercado-servicos-streaming.htm

[2] https://exame.com/tecnologia/spotify-continua-a-crescer-e-aumenta-numero-de-assinantes-em-21/

sistemas que dependem de notas ou *feedback* dos usuários, apresenta vantagens intrínsecas associadas ao seu modelo. Primeiramente, destaca-se a questão do *Cold Start*, problema comum em SRs, onde a ausência de dados suficientes sobre novos usuários ou itens pode dificultar a geração de recomendações precisas ((FORMIGA, 2014)). Com dados binários, tal desafio é mitigado, visto que a análise restringe-se à presença ou ausência de músicas em *playlists*, sem a necessidade de um volume expressivo de avaliações. O modelo binário tende a ser menos suscetível à esparsidade, uma vez que, em sistemas baseados em notas, muitos itens podem não possuir avaliações, resultando em matrizes de dados esparsas. No que tange à escalabilidade, a estrutura binária, por sua simplicidade, demonstra-se mais eficiente, especialmente quando se considera a expansão do número de usuários e itens ((SOUZA; MAILIDÚ, 2011)). Enquanto sistemas baseados em notas podem, por vezes, ser previsíveis em suas recomendações, a estrutura binária tem potencial para oferecer sugestões mais diversificadas e surpreendentes aos usuários. Além disso, ao evitar a dependência de notas, nosso modelo se distância de vieses de polarização que podem surgir de avaliações subjetivas.

No contexto deste trabalho de pesquisa, as medidas de similaridade desempenham um papel crucial na análise e compreensão dos dados binários utilizados para a construção do SR. Estas medidas são essenciais para avaliar a afinidade entre os usuários, um elemento central na geração de recomendações precisas e relevantes. Neste contexto, algumas das medidas de similaridade exploradas incluem a Similaridade de *Jaccard*, a Similaridade de *3W-Jaccard*, a Similaridade de *Sorensen-Dice*, a Similaridade de *Ochai*, a Similaridade de *Sokal & Sneath*, e a Similaridade de *Sokal & Michener*.

A Similaridade de *Jaccard* e a Similaridade de *3W-Jaccard* (RICCI; ROKACH; SHAPIRA, 2011) se destacam por sua capacidade de comparar a sobreposição de itens em *playlists*, sendo especialmente adequadas para dados binários onde a presença ou ausência de itens é o foco principal. A Similaridade de *Jaccard* compara a proporção de itens comuns entre as *playlists*, enquanto a Similaridade de *3W-Jaccard* atribui diferentes pesos às correspondências conjuntas, levando em consideração a importância dessas correspondências. Além disso, a Similaridade de *Sorensen-Dice* (VERMA; AGGARWAL, 2020) se revela valiosa na análise de dados binários, uma vez que considera tanto as correspondências conjuntas quanto a ausência de correspondência, penalizando a falta de correspondência. Esta medida é particularmente útil na identificação de padrões de sobreposição e na geração de recomendações precisas. A Similaridade de Ochai (WOLF, 1996), por sua vez, destaca-se ao atribuir pesos tanto às correspondências conjuntas quanto à ausência de correspondência, tornando-a adequada para avaliar a similaridade entre *play-*

*lists* onde a falta de um item é tão relevante quanto a sua presença. Por fim, as medidas de Similaridade de *Sokal & Sneath* (ROHLF; SOKAL, 1962) e a Similaridade de *Sokal & Michener* (ALBUQUERQUE et al., 2016), aplicáveis a dados binários, são eficazes na comparação de *playlists* e na identificação de padrões de consumo semelhantes. Elas consideram as correspondências conjuntas e a ausência de correspondência, fornecendo *insights* essenciais para a geração de recomendações personalizadas. Em síntese, as medidas de similaridade em um contexto de dados binários representam uma parte fundamental na construção de um Sistema de Recomendação eficiente. A escolha criteriosa entre essas medidas influencia diretamente a qualidade das recomendações geradas, assegurando que os usuários recebam sugestões alinhadas com seus gostos musicais e preferências.

Dessa maneira, será viável explorar as preferências musicais dos consumidores, identificando padrões de consumo e fornecendo recomendações personalizadas com base nas *playlists* criadas pelos consumidores. O desafio reside na utilização de dados binários, isto é, informações que indicam a presença ou ausência de uma música em uma *playlists*, para efetuar as recomendações com precisão e relevância. O SR binário proposto opera por meio de múltiplas etapas. O processo *Leave One Out* desempenha um papel crucial no desenvolvimento do SR. Sua aplicação precede a fase de construção do sistema, sendo destinado a determinar características essenciais. O objetivo primordial do *Leave One Out* é a especificação da medida de similaridade mais adequada e o melhor valor de $k$ a ser considerado no sistema. Esta etapa consiste na seleção aleatória de entradas 1 (indicando a presença de músicas) em *playlist* e sua transformação em 0 (indicando ausência) para possibilitar estimativas. A partir da análise dos resultados obtidos com diferentes medidas de similaridade e valores de $k$, será possível identificar as configurações que maximizam o desempenho do SR, contribuindo para recomendações mais precisas e relevantes.

Inicialmente na construção do SR, a similaridade entre as *playlists* é calculada com base em medidas de similaridades citadas a cima. Em seguida são identificadas as *playlists* mais similares, ou seja, vizinhos mais próximos. A estimação das recomendações é realizada por meio de médias ponderadas, em que a medida de similaridade atribui diferentes pesos à presença ou ausência de músicas nas *playlists* de referência, este cálculo é chamado de Chance de Consumo. As cinco estimativas mais elevadas são, então, sugeridas para serem adicionadas à *playlist* alvo. Esse sistema de recomendação binário confere a capacidade de apresentar sugestões musicais personalizadas e relevantes, superando os desafios inerentes aos dados binários, como a ausência de notas explícitas dos usuários, aprimorando, assim, a precisão e a qualidade da experiência de recomendação.

O objetivo geral deste trabalho consiste em desenvolver um Sistema de Recomendação, adaptado ao contexto de dados binários do *Spotify*. Esse sistema será construído por meio da análise comparativa de diferentes medidas de similaridade, centrando-se na Filtragem Colaborativa com ênfase na similaridade entre os usuários, ou seja, as *playlists* Os objetivos específicos são: compreender o funcionamento de um Sistema de Recomendação com dados binários; comparar as diferentes abordagens de medidas de similaridades como a *Jaccard*, *3W-Jaccard*, S*orensen-Dice*, *Ochai*, *Sokal & Sneath* e *Sokal & Michener* para dados binários, e avaliar o Sistema de Recomendação por Filtragem Colaborativa utilizando dados binários.

O trabalho será abordado em quatro capítulos. No Capítulo 2, foram apresentados a base dados com as suas informações e as técnicas estatísticas aplicadas nas análises. Já no Capítulo 3, teremos os resultados obtidos através dos dados analisados. Por fim, no Capítulo 4, serão detalhadas as principais conclusões alcançadas pela pesquisa.

# 2 Materiais e Métodos

Neste capítulo serão abordados os materiais e métodos para aplicação de um Sistema de Recomendação baseado em Filtragem Colaborativa.

## 2.1 Spotify

O *Spotify*, é um serviço digital em que os usuários têm acesso instantâneo a milhares de músicas, *podcasts*, vídeos e outros tipos de conteúdos[3]. Desenvolvido pelos empresários suecos Daniel Ek e Martin Lorentzon foi idealizado, em 2006, com a proposta de ser uma forte ferramenta de combate a pirataria na indústria musical. O lançamento ocorreu em 2008, após inúmeros testes, na expectativa de revolucionar o mercado fonográfico. Na época, o concorrente *iTunes*[4], que é um reprodutor de áudio, com cinco anos de mercado tinha um público restrito devido ao alto custo do serviço e a popularidade do *Torrent*[5], distribuidor de dados e grandes arquivos pela *Internet*, crescia. Somente em 2011, o serviço alcançou a América pois anteriormente era limitado a Europa. O sucesso da plataforma musical teve personalidades importantes que aprovaram o serviço, pode-se citar o criador do *Facebook* o Mark Zuckerberg. Após a expansão do *streaming*, em dois anos, houve a redução de 25% da pirataria de músicas na Suécia, local de origem da plataforma (BALADY, 2020). O serviço chegou no Brasil em 2014, já contava com a marca de 30 milhões de músicas, versão em aplicativo e *web player*, em *desktop*, *tabletes* e *smartphones* (BALADY, 2020).

Para manter seus consumidores cada vez mais engajados, o *Spotify*, através de Sistemas de Recomendação constrói uma *home* personalizada conforme as preferências do ouvinte. A título de exemplo, em junho de 2022 após o anúncio realizado pela cantora Beyoncé no *Instagram*[6] sobre o lançamento do próximo álbum o *Spotify* acrescentou reco-

[3]https://support.spotify.com/br/article/what-is-spotify/
[4]https://www.apple.com/br/itunes/
[5]https://www.utorrent.com/intl/pt$_b$r/
[6]https://www.instagram.com/beyonce/

mendações de álbuns da cantora para o público que consume os seus conteúdos (Figura 1). Um trilho personalizado, como este, precisa responder a seguinte pergunta: "O usuário consumiu algum item do artista ou não?" A resposta binária dada pelo histórico de consumo é uma ferramenta potente para ser usada em um momento de grande ansiedade para os fãs. O mecanismo de *marketing* utilizado pela plataforma através de um Sistema de Recomendação agrega valor e confiabilidade aos seus consumidores.

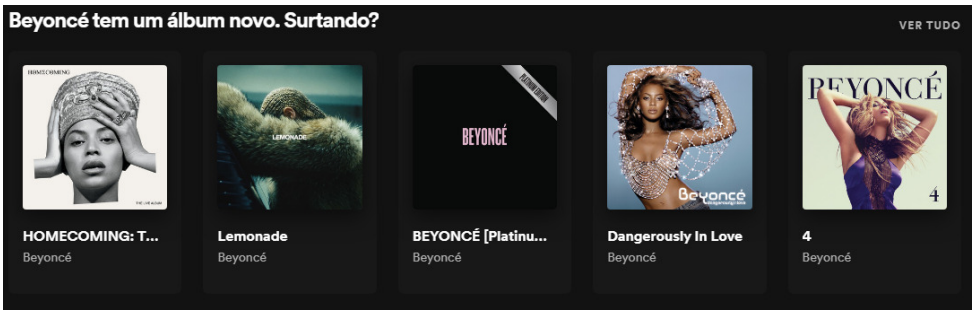


Figura 1: Trilho "Beyoncé tem um álbum novo. Surtando?" na *Home* do consumidor

O *Spotify* também traz para os usuários, recomendações que não estão ligadas diretamente a um artista. Pode-se observar na Figura 2, uma recomendação que considera a hora do dia para recomendar *playlists* voltadas para atividades matutinas.

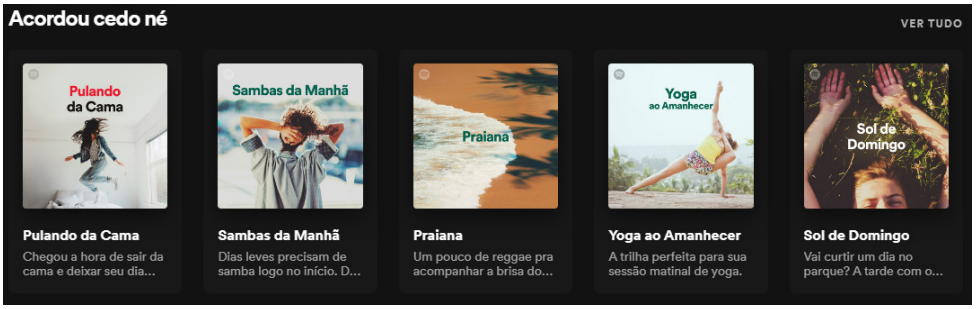


Figura 2: Trilho "Acordou cedo né?" na *Home* do consumidor

Esta variedade de recomendações personalizadas são características fortes do serviço. O comportamento destes tipos de usuários e o consumos feitos por eles, permitem sugerir através dos dados binários de consumo uma recomendação mais apropriada para o assinante.

## 2.2 Materiais

Para este trabalho, serão utilizados bases de dados da *RecSys Challenge* 2018. A *ACM Conference on Recommender Systems (RecSys)*, é o principal fórum internacional

para a apresentação de novos resultados de pesquisa, sistemas e técnicas no amplo campo dos Sistemas de Recomendação (SR). A *RecSys*[7] reúne os principais grupos internacionais de pesquisa que trabalham em SRs, com muitas das principais empresas mundiais ativas no comércio eletrônico e outros domínios adjacentes. Ao longo dos anos, tornou-se a conferência anual mais importante para a apresentação e discussão de pesquisas sobre SR.

O banco de dados conhecido como *Million Playlist Dataset* (MPD) foi cedido pela própria equipe do *Spotify* para a realização do desafio da *RecSys*. O desafio era realizar recomendação de músicas após o consumidor ouvir toda uma *playlist*, sendo assim, ao finalizar uma lista de reprodução a plataforma recomendaria músicas para estender o consumo para além do final das listas de reprodução existentes.

As informações da base de dados são resultados de *playlists* criadas pelos usuários que residem nos Estados Unidos, com ao menos 13 anos de idade, durante o período de 1º de Janeiro de 2010 e antes de 1º de Dezembro de 2017. As *playlists* eram públicas no momento em que o MPD foi gerado, com ao menos 5 músicas e não mais que 250 músicas, no mínimo 3 artistas únicos e 2 álbuns únicos, não possuem músicas locais que não estão no Spotify e com ao menos um seguidor. Na época, o *Spotify* continha mais de 4 bilhões de *playlists* públicas na plataforma, este conjunto de dados do MPD com 1 milhão de *playlists* consiste em mais de 2 milhões de faixas exclusivas de quase 300.000 artistas e representa o maior conjunto de dados públicos de *playlists* de música do mundo.

A base de dados empregada neste trabalho consiste em 869 observações, onde cada observação representa uma *playlist* única, e é composta por 12 variáveis. Vale ressaltar que as variáveis não descrevem informações de consumo ou avaliação das músicas, mas sim contém as músicas específicas dentro de cada *playlist*. Na Tabela 1 será apresentada as informações da base de dados com as variáveis sobre as *playlists*. Sendo assim, fornecendo informações cruciais para a compreensão das *playlists* e servindo como base para o desenvolvimento do Sistema de Recomendação proposto.

[7]https://recsys.acm.org/recsys22/

Tabela 1: Descrição das variáveis referente as *playlists* do *Spotify*

| *Variáveis originais* | Descrição |
|---|---|
| name | Nome |
| collaborative | Se é colaborativa |
| pid | Posição na base de dados |
| modified_at | Última modificação |
| num_tracks | Quantidade de músicas |
| num_albuns | Quantidade de álbuns |
| num_followers | Quantidade de seguidores |
| tracks | Informações das músicas |
| num_edits | Quantidade de edições |
| duration_ms | Duração em milissegundos |
| num_artists | Quantidade de artistas |
| description | Descrição feita pelo o consumidor |

A variável *tracks*, é composta por uma lista de informações das músicas presentes em cada *playlist*. A seguir será abordado quais as informações são vistas em *tracks*:

Tabela 2: Informações da variável *track*

| *Variáveis originais* | Descrição |
|---|---|
| pos | Posição da música na *playlist* |
| artist_name | Nome do Artista |
| track_uri | Código da música no *Spotify* |
| artist_uri | Código do artista no *Spotify* |
| track_name | Nome da música |
| album_uri | Código do álbum no *Spotify* |
| durations_ms | Duração da música em milissegundos |
| album_name | Nome do álbum |

Essas informações estão presentes em todas as *playlists* da base de dados que usaremos ao longo do trabalho.

## 2.3 Sistemas de Recomendação

Os consumidores de plataformas digitais, encontram uma imensidão de produtos e conteúdos sendo ofertados a todo momento tornando cada vez mais difícil a tarefa de encontrar algo que lhe agrade (ROZENDO, 2017). Os Sistemas de Recomendação (SR), são caraterizados por uma família de algoritmos que são capazes de filtrar informações úteis de consumo para sugerir itens personalizados aos clientes (LIMA, 2019). Ao analisar os dados de consumo é feita a extração das informações necessárias para realizar as predições, ou seja, através de um conjunto de dados com informações de Usuários e Itens os algori-

timos de um SR geram recomendações para os consumidores (SOUZA; MAILIDÚ, 2011). Este processo de recomendação é descrito na Figura 3.


Figura 3: Processo de Recomendação

Sejam $U$ o conjunto de todos usuários presentes na base de dados, $I$ o conjunto de todos os itens disponíveis e $I_u$ o conjunto dos itens que não foram consumidos ou avaliados pelo usuário $u$. O Sistema de Recomendação SR irá predizer quais itens o usuário $u$ teria maior preferência de consumo para cada um dos itens não consumidos em $I_u$.

Os Sistemas de Recomendação lidam com o problema de *Cold Start*, definido como a ausência ou pouco volume de dados iniciais que prejudicam o processo de recomendação. Na Recomendação Baseada em Conteúdo, a entrada de novos itens na base de dados será usada, suas características para recomendar outros itens com pelo menos 1 atributo igual ao item consumido. Sendo assim, o *Cold Start* não irá prejudicar as sugestões, pois utiliza aspectos do item. A recomendação Baseada em Filtros Colaborativos, precisa de informações como o consumo ou não de um item, por exemplo, para calcular a similaridade entre os consumidores (FORMIGA, 2014).

No contexto deste trabalho, que se concentra na Recomendação Baseada em Filtros Colaborativos com dados binários, a preocupação com o *Cold Start* não se aplica, uma vez que não faremos uso de avaliações ou feedbacks dos consumidores para as recomendações (MASSON, 2016). Em nosso cenário, estamos lidando com dados que indicam a presença ou ausência de uma música em uma *playlist* específica, o que torna a ausência ou um volume inicial limitado de dados irrelevantes para o SR. Dessa forma, o *Cold Start* não representa uma preocupação ou desafio a ser superado.

### 2.3.1 Recomendação Baseada em Conteúdo

Segundo SILVA (2019), a Recomendação Baseada em Conteúdo (*Content-Based Recommendation*, do inglês) tem a finalidade de recomendar baseado na similaridade física entre itens ou entre usuários.

Na Figura 4, temos um Sistema de Recomendação Baseada em Conteúdo por Usuários. Desta forma, as características similares são dos usuários são essenciais para realizar a recomendação. Neste exemplo, o Consumidor A ouviu a música *Descobridor dos Sete Mares* e o Consumidor B que tem as mesmas características como faixa etária, sexo e também nacionalidade irá receber este item como uma recomendação.

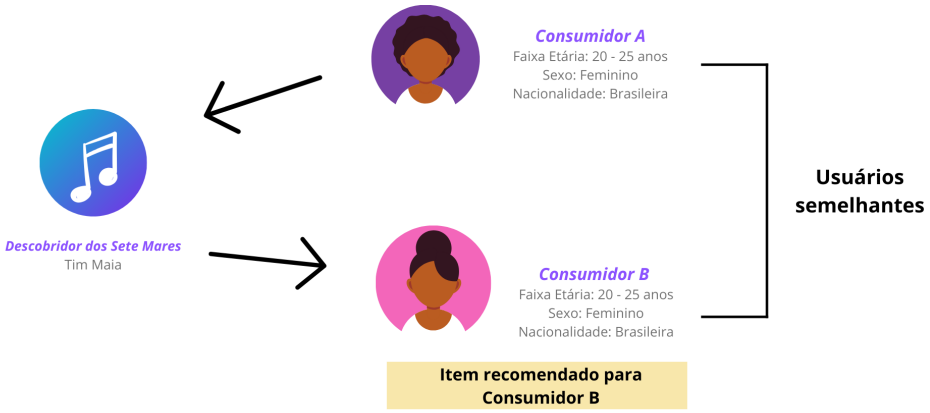


Figura 4: Recomendação Baseada em Conteúdo por Usuários

Na Recomendação Baseada em Conteúdo por Item, Figura 5, nota-se que o Consumidor ouviu a canção *Don't Hurt Yourself*, a filtragem busca uma música não consumida por ele, porém que contém as características em comum ao item ouvido. Observar-se que ambas as músicas são da cultura negra, sendo assim, a filtragem buscou por essa propriedade igual. Por consequência, a música *We Cry Together* é recomendada ao consumidor-alvo.


Figura 5: Recomendação Baseada em Conteúdo por Item

### 2.3.2 Recomendação Baseada em Filtros Colaborativos ou Cosumo

Na Recomendação Baseada em Filtros Colaborativos ou Cosumo (*Collaborative-Filtering*, do inglês), encontrará usuários ou itens similares conforme o padrão de consumo dos usuários. Sendo assim, a recomendação será feita para usuários ou itens com perfis semelhantes de consumo. Portanto, dois usuários serão semelhantes se eles consomem itens em comum, ou, dois itens serão considerados semelhantes se eles são consumidos por usuários em comum (SILVA, 2019).

Na Figura 6, o Consumidor A e o Consumidor B ouviram as canções *Crazy in Love* e *No One*, nota-se que o perfil de consumo deles são similares. Somente o consumidor A ouviu *Redbone*, desta forma a música pode ser uma recomendação para o consumidor B considerando que eles têm consumo similares.


Figura 6: Recomendação Baseada em Filtros Colaborativos por Usuário

No exemplo da Figura 7, os itens *Heated* e *Happy* são consumidos por usuários em comum. O consumo do item *Heated* feito pelo o Consumidor C desencadeia em uma recomendação do item *Happy* pois eles são cosumidos similarmente pelos consumidores A e B.

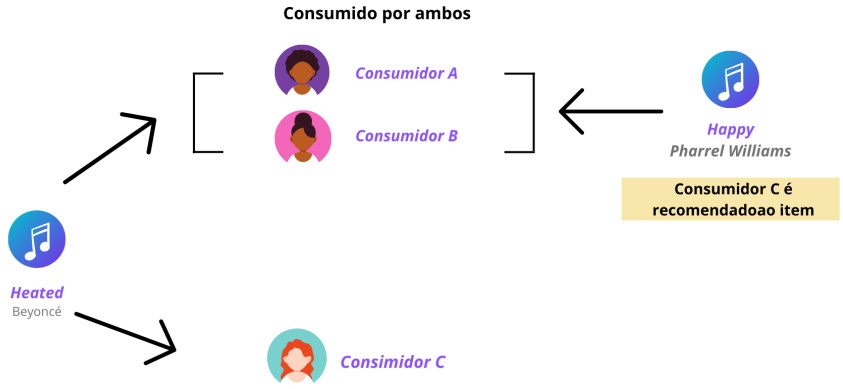


Figura 7: Recomendação Baseada em Filtros Colaborativos por Item

### 2.3.3 Recomendação Híbrida

A Recomendação Híbrida, utiliza diferentes tipos algoritmos de filtragem como vimos anteriormente para a recomendação. Em algumas situações, são necessários aplicar mais de um algoritmo para obter os melhores resultados. Desde modo, a recomendação tem

alta complexidade de implementação (SILVA, 2019).

## 2.4 Filtragem Colaborativa

A Filtragem Colaborativa visa recomendar itens. A predição forma um conjunto de consumidores com interesses em comuns como gosto musical, por exemplo, para realizar sugestões de músicas que o consumidor-alvo ainda não ouviu ou desconhece a partir do outro consumidor similar. Os interesses destes consumidores são mensurados de forma explicita como avaliação das músicas ou implícita tal como o consumo ou não da música, por exemplo (GODINHO; VASCONCELOS, ). Uma abordagem comumente utilizada na filtragem colaborativa é a Matriz de Utilidade, que representa a relação entre usuários e itens em uma forma tabular. Essa matriz contém informações sobre o consumo dos usuários em relação aos itens, como músicas ou *playlists*, e é preenchida com valores que refletem a utilidade percebida pelos usuários para cada item. A Matriz de Utilidade é construída a partir dos dados coletados, que podem incluir classificações explícitas, como avaliações numéricas, ou informações implícitas, como histórico de reprodução ou presença de um item em uma *playlist*. Com base nesses dados, algoritmos de recomendação podem analisar a matriz e identificar padrões, relações de similaridade entre usuários e itens, e gerar recomendações personalizadas com base nesses *insights*. A matriz de utilidade é uma representação fundamental no processo de filtragem colaborativa, permitindo a análise e exploração dos dados para oferecer recomendações relevantes aos usuários. A Filtragem Colaborativa pode ser dividida em dois métodos:

i) **Método Baseado na Vizinhança**: conhecido também como Método Baseado em Memória. Esta abordagem utiliza a matriz de similaridade para encontrar itens ou usuários semelhantes para realizar a recomendação (SILVA, 2019). Nas Seções 2.4.3 e 2.4.4 seguintes serão abordados respectivamente estes conceitos de Matriz de Utilidade e Matriz de Similaridade. Neste trabalho é utilizado o Método Baseado na Vizinhança.

ii) **Método Baseado em Modelos**, através dos modelos de Machine Learning será feita previsões do consumo do usuário para cada item. Sendo assim, será extraído alguns dados da base para criação de um modelo capaz de realizar a recomendação sem a necessidade de usar todas observações da base, ou seja, é usado uma amostra da base de dados. Este tipo de abordagem tem como pontos fortes a velocidade e escalabilidade (SILVA, 2019).

O Método Baseado na Vizinhança depende da similaridade entre usuários ou itens

específicos, o Método Baseado em Modelos se baseia na construção de modelos preditivos mais amplos para fazer recomendações. Cada abordagem tem suas vantagens e desvantagens, e a escolha depende do contexto específico e dos requisitos do sistema de recomendação. As desvantagens destes sistemas são esparsidade dos dados e a escalabilidade, que é capacidade de manipular grande volumes de dados. A esparsidade dos dados está presente nos Métodos de Vizinhança, pois precisam de correspondências exatas, isto afeta a precisão dos SR. Uma vez definido o relacionamento entre usuários que consumiram pelo menos dois itens em comum, muitos "pares"de consumidores não terão nenhuma associação entre eles devido à imensidão de itens. Os consumidores ativos do *Spotify* que ouviram músicas e *podcasts* possivelmente não interagiram nem com 1% dos itens disponíveis na plataforma (SOUZA; MAILIDÚ, 2011).

A escalabilidade está relacionada ao crescimento do número de consumidores e itens, os algoritmos de vizinhança demandam um esforço computacional para acompanhar a evolução da base de dados e os SR acabam enfrentando problemas de escalabilidade (SOUZA; MAILIDÚ, 2011).

### 2.4.1 Avaliação implícita e explicita

Existem diversas maneiras do usuário demonstrar interesse ou predileção aos itens consumidos. O *feedback* explícito é obtido por dados diretos, ou seja, dados de interação intencional do consumidor quando ele interage diretamente na avaliação de um item dando uma nota, indicando a satisfação por "gostei ou não gostei", exemplo de avaliação binária, ou por um sistema de estrelas de 0 a 5. Já no *feedback* implícito, os dados são obtidos indiretamente, que pode ser mensurados como a frequência que o consumidor ouve uma música, tempo de música ouvida ou até mesmo se ouviu o item (AZAMBUJA; MORAIS; FILIPE, 2021).

As informações coletadas de cada usuário com os *feedbacks* dos itens irão preencher a Matriz de Utilidade para gerar as recomendações. Neste trabalho a base de dados, descrita na Seção 2.2, contém informações de feedback implícito, sendo apenas a informação se o item foi consumido ou não. Mais detalhes sobre a Matriz de Utilidade na Seção 2.4.3.

### 2.4.2 Dados Binários

O Sistema de Recomendação baseado em Filtragem Colaborativa, em geral, utiliza avaliações de itens dadas pelos usuários. Esta avaliação pode ser feita através da escala

Likert (LIKERT, 1932), com as notas 5, 4, 3, 2 e 1, em que respectivamente indicam, "Adorei", "Gostei", "Indiferente", "Não gostei" e "Detestei". Uma abordagem comum é a utilização de dados de classificação ou avaliação, em que os usuários expressam suas preferências atribuindo uma pontuação ou avaliação aos itens. Esses dados podem ser representados em uma matriz de avaliação, em que cada linha representa um usuário, cada coluna representa um item e os valores indicam a avaliação dada pelo usuário ao item.

Em cenários que não se tem a avaliação do consumidor, os dados binários são utilizados como forma de solucionar esta situação para que possa ser feita a recomendação. O que acontece com os dados do *Spotify* que iremos utilizar. Essa representação binária possibilita que a análise, permitindo a identificação de padrões de consumo entre os usuários. Além disso, dados binários também podem ser utilizados para representar preferências implícitas, em que a presença ou ausência de interação do usuário com um item é considerada como uma indicação de interesse ou desinteresse. Independentemente da forma de dados utilizada, é necessário aplicar técnicas e algoritmos de Filtragem Colaborativa adequados para extrair informações úteis desses dados, identificar similaridades entre usuários e itens, e fornecer recomendações personalizadas e relevantes para cada usuário.

Neste trabalho, será utilizado dados binários. Assim sendo, as transações entre usuários e itens serão definidos por 0 e 1. A entrada 0 é tratada como o não consumo do item por não conhecê-lo ou o usuário conhece o item mas optou por não consumir. Os dados binários, neste cenário as entradas 0 não são definidas como o não desejo pelo item. Tendo como exemplo, o caso em que uma música não pertence a uma *playlist* ela será entendida como o não consumo da música pela *playlist* ou apesar de ser conhecida ela não está na *playlist*. A entrada 1 corresponde pelo consumo do item pelo usuário, ou seja, se uma música pertencer a uma *playlist*, esta transação é compreendia pelo consumo da música pela *playlist*.

Os dados binários aplicados a um Sistema de Recomendação (SR) geram algumas limitações na hora de recomendar. Limitações como não tem uma classificação para o item indicando se ele é relevante ou não para o usuário, isto é, se uma música que não foi inserida na *playlist* propositalmente por não ser relevante ou porque quem fez a *playlist* não a conhecia. Caso, esta música não tenha sido inserida por não ser relevante e o SR sugerir ela pode tirar a credibilidade do sistema, por exemplo. Além disso, não é possível utilizar as métricas como *Recall, Precison* e *Medida F* para verificar a qualidade do sistema. As métricas de *Recall*, *Precisão* e *Medida F* são indicadores essenciais para

avaliar o desempenho de sistemas de recomendação. O *Recall* mensura a proporção de itens relevantes que foram corretamente identificados, a *Precisão* avalia a precisão das recomendações feitas, enquanto a *Medida F* combina essas métricas, fornecendo uma visão balanceada entre ambas. Em conjunto, essas métricas permitem uma avaliação abrangente do quão eficaz é o sistema em fornecer recomendações relevantes, sendo cruciais para otimizar a experiência do usuário em plataformas de recomendação. Portanto, a construção do SR deve considerar todos esses fatores para evitar eventuais problemas na recomendação.

### 2.4.3 Matriz de Utilidade

A Matriz de Utilidade pretende organizar as informações da base de dados segundo a preferência consumo dos itens pelos indivíduos. A dimensão é dada por $N \times M$, sendo $N$ o número de usuários na base de dados e $M$ o número de itens. Cada vetor-linha é a representação de um consumidor e cada vetor-coluna irá representar o item. Isto posto, a $i$-ésima linha contém dados da avaliação (explícito ou implícito) do consumidor $i$ e a $j$-ésima coluna apresenta dados das avaliações recebidas pelo item $j$. Segundo a Matriz 2.1, cada célula $a_{i,j}$ da Matriz de Utilidade é composta pela avaliação dada pelo consumidor $i$ para o item $j$, para $i \in \{1, ..., N\}$ e para $j \in \{1, ..., M\}$ (GOMES et al., 2019). Para este trabalho, a Matriz de Utilidade é composta pela informação se uma música, por exemplo, é consumida ou não em uma determinada *playlist*, na Seção 2.2 é apresentada a base de dados. Logo, as entradas serão do tipo binário, isto é, 0 - se é não consumido pela *playlist* e 1 - se é consumido pela *playlist*, por exemplo.

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,M} \\ a_{2,1} & a_{2,2} & \dots & a_{2,M} \\ \dots & \dots & \ddots & \dots \\ a_{N,1} & a_{N,2} & \dots & a_{N,M} \end{bmatrix} \tag{2.1}$$

Através da Matriz de Utilidade é possível realizar diversas formas de recomendações usando a base de dados. Neste trabalho, os dados são referentes à artistas, músicas e *playlists* que serão usados na recomendação. Cada uma destas possibilidades geram uma Matriz de Utilidade diferente. O exemplo a seguir abordam uma forma de como construir a Matriz de Utilidade segundo o que será recomendado.

**Exemplo 2.3.1** *Considere uma Matriz de Utilidade que é composta por 2 playlists (Usuários) e 7 músicas (Itens). O objetivo é recomendar uma música ao consumidor com*

*base nas playlists ouvidas por ele.*

Na Tabela 3, as *playlists* A e B da nossa base de dados. Estas *playlists* assumem o papel de Usuários na Matriz de Utilidade, sendo assim, serão as elas que receberam as recomendações. Já as músicas são os Itens presentes ou não em cada uma das *playlists*. Caso a música pertença a *playlists* ela é rotulada como *Pert.* e se não estiver é *Não Pert.*.

Tabela 3: Matriz de Utilidade

| *Playlist* | Música 1 | Música 2 | Música 3 | Música 4 | Música 5 | Música 6 | Música 7 |
|---|---|---|---|---|---|---|---|
| A | Pert. | Pert. | Não pert. | Pert. | Pert. | Não pert. | Pert. |
| B | Não pert. | Pert. | Não pert. | Pert. | Não pert. | Pert. | Pert. |

A Matriz de Utilidade 2.2 é construída com dados binários, as entradas 0 representam que a música não está presente na *playlist* correspondente e 1 quando a música está presente.

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \tag{2.2}$$

### 2.4.4 Matrizes de Similaridade

Ao longo desta seção, serão apresentadas as medidas de similaridade e a partir destes resultados é realizada a recomendação baseada em Filtragem Colaborativa.

A Matriz de Similaridade entre Usuários, apresentada na Equação 2.3, é definida por uma matriz quadrada $N \times N$, em que $N$ é o número total de usuários e as entradas $S^U_{i,k}$ é a representação da similaridade entre o usuário $i$ com o usuário $k$. No cálculo de Similaridade entre Usuários, cada usuário é um vetor no $R_n$ corresponde pelo vetor-linha da Matriz de Utilidade ($a_i$). Neste trabalho usaremos a Matriz de Similaridade entre Usuários.

$$\begin{bmatrix} S^U_{1,1} & S^U_{1,2} & \dots & S^U_{1,N} \\ S^U_{2,1} & S^U_{2,2} & \dots & S^U_{2,N} \\ \dots & \dots & \ddots & \dots \\ S^U_{N,1} & S^U_{N,1} & \dots & S^U_{N,N} \end{bmatrix} \tag{2.3}$$

A Matriz de Similaridade entre Itens, Equação 2.4, uma matriz quadrada $M \times M$, em que cada entrada será $S^I_{j,k}$ representando a similaridade entre o item $j$ com o item $k$ e $M$ é o número de itens disponíveis para consumo.

$$\begin{bmatrix} S^I_{1,1} & S^I_{1,2} & \dots & S^I_{1,M} \\ S^I_{2,1} & S^I_{2,2} & \dots & S^I_{2,M} \\ \dots & \dots & \ddots & \dots \\ S^I_{N,1} & S^I_{N,1} & \dots & S^I_{M,M} \end{bmatrix} \tag{2.4}$$

Pode-se perceber que a Matriz de Similaridade entre Itens e a Matriz de Similaridade dentre Usuários são semelhantes na estrutura matricial, ou seja, são matrizes simétricas já que $S^U_{i,k} = S^U_{k,i}$ e também $S^I_{j,k} = S^I_{k,j}$ $\forall j, k$ (GOMES et al., 2019).

A mensuração da semelhança entre usuários e entre itens pode-se ser feita por diferentes medidas de similaridade. Para este trabalho, foram escolhidas as diferentes abordagens de medidas de similaridades como a *Jaccard*, *3W-Jaccard*, *Sorensen-Dice*, *Ochai*, *Sokal & Sneath* e *Sokal & Michener* para dados binários. Essas similaridades serão utilizadas para encontrar as *playlists* similares e iremos escolher aquela que teve um desempenho deestaque para ser incorporada ao Sistema de Recomendação.

#### 2.4.4.1 Recomendação a partir da Similaridade entre Usuários

Com Matriz de Similaridade entre Usuários, será selecionado os **k** usuários mais próximos do usuário $i$ que cosumiram o item $j$. Esta lista de usuários mais próximos é denotada de $G^U_{i,j}$, sendo assim, teremos as *playlists* mas similares para usar no Sistema de Recomendação. O consumo realizado pelo usuário $i$ ao item $j$ será prevista utilizando a média ponderada dos consumos realizados pelos usuários da lista $G^U_{i,j}$ ao item $j$, a ponderação é calculada conforme a medida de similaridade escolhida, na qual tem-se usuários mais próximos ao usuário $i$ ganharão pesos maiores. Esta previsão do consumo do item $j$ pelo usuário $i$ é representada por $\hat{a}^U_{i,j}$,

$$\hat{a}^U_{i,j} = \frac{\sum_{\{\forall k \in G^U_{i,j}\}} a_{k,j} \cdot s^U_{i,k}}{\sum_{\{\forall k \in G^U_{i,j}\}} s^U_{i,k}} \tag{2.5}$$

#### 2.4.4.2 Recomendação a partir da Similaridade entre Itens

De forma análoga, a Similaridade entre Itens irá prever o consumo do usuário $i$ ao item $j$. A partir da Matriz de Similaridade entre Itens serão selecionados os **n** itens mais semelhantes ao item $j$ que foram consumindos pelo usuário $i$, esta lista de usuários será denominada de $G^I_{i,j}$. Portanto, o consumo realizado pelo usuários $i$ ao item $j$ será prevista

com base na média ponderada dos consumos aos itens similares ao item $j$, a ponderação é feita segundo a similaridade definida, na qual os itens mais próximos ao item $j$ receberão maiores pesos. A previsão é representada por $\hat{a}^I_{i,j}$,

$$\hat{a}^I_{i,j} = \frac{\sum_{\{\forall k \in G^I_{i,j}\}} a_{k,j} \cdot s^I_{i,k}}{\sum_{\{\forall k \in G^I_{i,j}\}} s^I_{i,k}} \tag{2.6}$$

Neste trabalho, será aplicado a Recomendação a partir da Similaridade entre Usuários, sendo assim, as *playlists* (Usuários) receberão recomendações de múscias (Itens).

### 2.4.5 Medidas de Similaridades

Existem diferentes medidas de similaridade que são utilizadas nos Sistemas de Recomendação para avaliar a similaridade entre usuários e itens. Neste trabalho, serão abordadas as similaridades mais adequadas para os dados binários (RAMÍREZ; BATYRSHIN, 2016). Nas próximas Subseções serão tratadas as Similaridades de *Jaccard* e *3W-Jaccard*, *Sorensen-Dice*, *Ochai*, S*okal & Sneath* e *Sokal & Michener*.

Para melhor compreensão, ao longo das subseções serão apresentados exmplos para cada similaridade apresentada. Os dados serão extraídos da Matriz de Utilidade a seguir:

Tabela 4: Matriz de Utilidade

| *Playlist* | Música 1 | Música 2 | Música 3 | Música 4 | Música 5 | Música 6 | Música 7 |
|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| B | 0 | 1 | 0 | 1 | 0 | 1 | 1 |

#### 2.4.5.1 Similaridade de Jaccard e 3W-Jaccard

A Similaridade de *Jaccard*, também é conhecido como Índice de *Jaccard*, mede a similaridade e a diversidade dos conjuntos de amostras (RICCI; ROKACH; SHAPIRA, 2011). Para este trabalho, por exemplo, o cálculo de similaridade de duas *playlists I, J* pelo coeficiente *Jaccard* compara a soma das músicas em comum sobre a soma das músicas que não são repetidas nas *playlists*. A repersentação em operações de conjuntos conjuntos da Similaridade de Jaccard é dada por:

$$S_{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.7}$$

Na qual A e B são dois conjuntos de itens, no numerador é feita a cardinalidade

da interseção e no denominador cardinalidade da união dos elementos. O resultado do coeficiente *Jaccard* é uma medida de similaridade que varia entre 0 e 1. Sendo que 1 indica que compartilham a maioria de seus elementos e 0 quando os elementos dos conjuntos são distintos (SOUSA; FERREIRA, 2018). No caso das duas *playlists*, se o resultado fosse 1 as *playlists* teriam as mesmas músicas. Caso o resultado fosse 0, as músicas presentes na *playlists* seriam todas diferentes.

**Exemplo 2.4.1** *Utilizando o Coeficiente de Jaccard, iremos calcular a similaridade entre a Playlist A e a Playlist B com os dados da Tabela 4, onde 1 indica o cosumo da música pela playlist e 0 indica que a música não foi consumida pela playlist.* Com base nos dados da Tabela 4 e na Equação 2.10, tem-se:

$$S_{Jaccard}(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.8}$$

$$= \frac{3}{6} = 0.5$$

As *Playlist A* e a *Playlist B* segundo o coeficiente de *Jaccard* é 0.5 indicando que existe similaridade entre elas.

A similaridade de *Jaccard* pode ser reescrita com o peso maior para o número de músicas consumidas em comum nas *playlists* e é denominada de *3W-Jaccard*. Assim como a Similaridade de *Jaccard*, a Similaridade de *3W-Jaccard* pode ser representada em conjuntos:

$$S_{3W-Jaccard}(A,B) = \frac{3 \cdot |A \cap B|}{3 \cdot |A \cap B| + |\overline{A} \cap B| + |A \cap \overline{B}|} \tag{2.9}$$

**Exemplo 2.4.2** *Para o conjunto de dados da Tabela 4, usaremos o Coeficiente de 3W-Jaccard para calcular a similaridade entre a Playlist A e a Playlist B. Pela Equação* **??**, *observa-se:*

$$S_{3W-Jaccard}(A,B) = \frac{3 \cdot |A \cap B|}{3 \cdot |A \cap B| + |\overline{A} \cap B| + |A \cap \overline{B}|} \tag{2.10}$$

$$= \frac{3 \cdot 3}{3 \cdot 3 + 2 + 1} = 0.75$$

Ao aumentar o peso dar um peso maior para o número de músicas consumidas em

comum em ambas as *playlists* pode-se notar um aumento na similaridade entre elas. O coeficiente para este exemplo indicou alta similaridade com 0.75.

#### 2.4.5.2 Similaridade de Sorensen-Dice

O Coeficiente de *Sorensen-Dice*, é uma medida amplamente utilizada para analisar a semelhança entre duas amostras de dados binários (VERMA; AGGARWAL, 2020). A similaridade encontrada será um resultado entre 0 e 1, onde o valor mais próximo de 1 apresenta maior grau de semelhança entre as *playlists*. A representção em conjuntos é dada por:

$$S_{Sorensen-Dice}(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{2.11}$$

Conforme a Equação 2.12, percebemos que as entradas nulas são excluídas do cálculo, e as entradas unitárias são duplicadas. Além disso, em comparação com o Coeficiente de *Jaccard*, que apresenta uma estrutura bem semelhante, ela terá um valor de similaridade maior devido ao peso 2 no número de músicas presentes em ambas as *playlists*.

**Exemplo 2.4.3** *Através da Tabela 4, iremos aplicar o Coeficiente de Sorensen-Dice para calcular a similaridade entre a Playlist A e a Playlist B. A Equação 2.12, será utilizada:*

$$S_{Sorensen-Dice}(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{2.12}$$

$$= \frac{2 \cdot 3}{2 \cdot 6} = 0.6667$$

Como dito anteriormente, o Coeficiente de *Sorensen-Dice* apesenta melhor desempenho que o Coeficiente de *Jaccard* porque as músicas em comum nas *playlists* recebem um peso maior. Logo, usando a mesma Matriz e Utilidade o Coeficiente de *Sorensen-Dice* apresentou uma similaridade de 0.6667 e foi maior que do Coeficiente de *Jaccard* que foi de 0.5. Portanto, podemos afirmar que a *Playlist* A e *Playlist* B apresentam uma alta similaridade.

#### 2.4.5.3 Similaridade de Ochai

O Índice de *Ochai* ou Similaridade de *Ochai* é baseada em dados binários de presença ou ausência de uma determinada característica. É muito conhecida e utilizada na área de botânica, ecologia e zoologia pois visa analisar as diferenças partilhadas entre algumas espécies de plantas ou organismo, por exemplo. A similaridade de *Ochai* é a forma binária da Similaridade de Cosseno (WOLF, 1996) e é dada pela seguinte representação em conjunto:

$$S_{Ochai}(A,B) = \frac{|A \cap B|}{\sqrt{(|A \cap B| + |\overline{A} \cap B|) \cdot (|A \cap B| + |A \cap \overline{B}|)}} \tag{2.13}$$

**Exemplo 2.4.4** *Através da Tabela 4, iremos aplicar o Coeficiente de Ochai para calcular a similaridade entre a Playlist A e a Playlist B. A Equação 2.4.5.3, será utilizada:*

$$S_{Ochai}(A,B) = \frac{|A \cap B|}{\sqrt{(|A \cap B| + |\overline{A} \cap B|) \cdot (|A \cap B| + |A \cap \overline{B}|)}} \tag{2.14}$$

$$= \frac{3}{\sqrt{5 \cdot 4}} = 0.6708$$

Para o exemplo em questão, a Similaridade de *Ochai* apresentou um resultado de 0.67 indicando que as *Playlist* A e *Playlist* B são semelhantes entre si.

#### 2.4.5.4 Similaridade de Sokal & Sneath

Assim como na Similaridade de *Ochai*(2.4.5.3), a Similaridade de *Similaridade de Sokal & Sneath* tem uma aplicação muito usual na ecologia. Esta medida foi desenvolvida para quantificar o procedimento de taxonomia e visa estimar os graus de semelhança entre espécies ou afinidade entre organismos (ROHLF; SOKAL, 1962). Para este trabalho, vamos usar do ponto de vista de similaridade entre *playlists* e a representação em conjuntos é:

$$S_{Sokal\&Sneath}(A,B) = \frac{|A \cap B|}{|A \cap B| + 2 \cdot |\overline{A} \cap B| + 2 \cdot |A \cap \overline{B}|} \tag{2.15}$$

Esta similaridade atribui peso duplo a não correspondências entre as *playlists*, ou seja, músicas que só aparecem em uma *playlist* e não na outra.

**Exemplo 2.4.4** *Através da Tabela 4, iremos aplicar o Coeficiente de Sokal & Sneath para calcular a similaridade entre a Playlist A e a Playlist B. A Equação 2.16, será utilizada:*

$$S_{Sokal\&Sneath}(A,B) = \frac{|A \cap B|}{|A \cap B| + 2 \cdot |\overline{A} \cap B| + 2 \cdot |A \cap \overline{B}|} \tag{2.16}$$

$$= \frac{3}{9} = 0.3333$$

Pode-se perceber que ao dar peso maior para as não correspondencias a medida de similaridade tende a ser menor. O valor encontrado pela similadridade de *Similaridade de Sokal & Sneath* foi de 0.333.

#### 2.4.5.5 Similaridade de Sokal & Michener

O coeficiente de Sokal & Michener também é conhecido como Concordância Simples *(Simple Matching)*. Esta medida correlaciona a similaridade entre amostras de dados binários. O valor da similaridade situa-se sempre entre 0 e 1, o maior valor apresenta-se elevado semelhanças entre as amostras de dados ALBUQUERQUE et al.. A representação em conjuntos é definida por:

$$S_{Sokal\&Michener}(A,B) = \frac{|A \cap B| + |\overline{A} \cap \overline{B}|}{|U|} \tag{2.17}$$

Portanto, as ausências conjuntas são incluídas no cálculo.

**Exemplo 2.4.5** *Através da Tabela 4, iremos aplicar o Coeficiente de Sokal & Michener para calcular a similaridade entre a Playlist A e a Playlist B. A Equação 2.18, será utilizada:*

$$S_{Sokal\&Michener}(A,B) = \frac{|A \cap B| + |\overline{A} \cap \overline{B}|}{|U|} \tag{2.18}$$

$$= \frac{3+1}{7} = 0.5714$$

### 2.4.6 Recomendação

O Sistema de Recomendação baseado em Filtragem Colaborativa entre Usuários proposta neste trabalho é abordada na Figura 8.

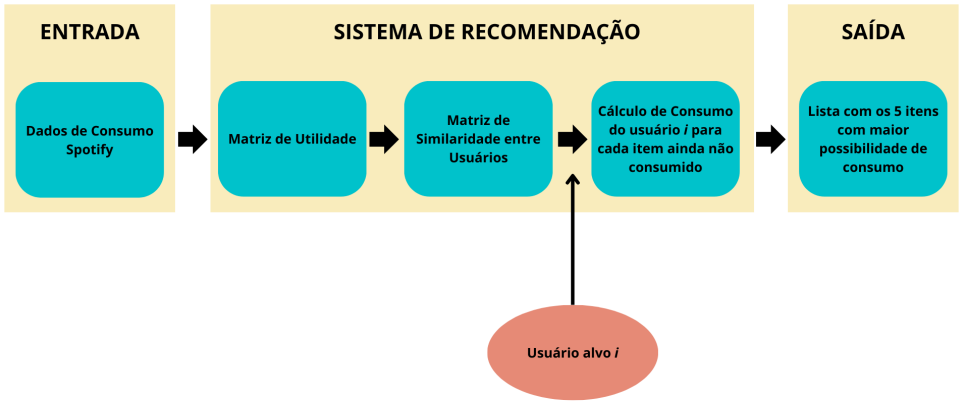


Figura 8: Fluxograma do Sistema de Recomendação

Na entrada, temos a base de dados referente ao consumo de músicas pelas *playlists*, ou seja, nas linhas teremos todas as *playlists* e nas colunas todas as músicas presentes na base de dados. O Sistema de Recomendação consiste em três etapas: a criação da Matriz de Utilidade na qual teremos as entradas 0 e 1, representando a ausência ou não de música na *playlist*, respectivamente. A partir desta matriz, é desenvolvida a Matriz de Similaridade entre Usuários com a medida de melhor desempenho utilizando $k$ *playlists* similares. A escolha da medida de similaridade e o $k$ ótimo serão abordados na Seção 2.4.6.1. Com a escolha da *playlist* que irá receber a recomendação, é realizado o cálculo do Consumo, que foi visto na Seção 2.4.4.1. Sendo assim, será feita a média ponderada entre a Matriz de Utilidade e a Matriz de Similaridade para obter o resultado do Consumo.

Com estes resultados do Consumo é possível realizar a recomendação de músicas para esta *playlist*. No *Spotify*, é definido que as *playlists* recebam 5 músicas recomendadas. Devido a este padrão, foi adotado que as *playlists* receberiam as 5 músicas com maior chance de consumo.

#### 2.4.6.1 *Leave-One-Out*

A técnica de Leave-One-Out, inserida na categoria de *Cross-Validation*, desempenha um papel crucial na determinação dos hiperparâmetros do Sistema de Recomendação

(SR) adotado neste estudo, que lida com dados binários (MONARD; BARANAUSKAS, 2003). O objetivo principal é que através desta técnica possamos escolher da medida de similaridade e do número de $k$ *playlists* mais similares a serem consideradas durante o processo de recomendação.

O processo começa tratando o consumo, representado por entradas iguais a 1 na Matriz de Utilidade, como desconhecido. Essa modificação é realizada retirando aleatoriamente duas entradas 1 de *playlist* por *playlist* por vez e substituindo-as por 0, permitindo a estimativa da entrada desconhecida.

Em seguida, efetua-se o cálculo de similaridade entre usuários utilizando diversas as medidas apresentadas. O resultado é utilizado de maneira comparativa para avaliar o desempenho de cada medida de similaridade. Além disso, as diferentes medidas são comparadas entre si, variando o valor de $k$.

Ao término desse processo, são obtidos diversos resultados para cada medida de similaridade com diferentes valores de $k$. Para defrinir a medida de similaridade que apresentou seu melhor desempenho quando combinada com o valor ótimo de k, os resultados serão plotados em *boxplots* para realizar a comparação. Sendo assim, cada medida de similaridade é plotada com suas diferentes $k$ *playlists*.

Através da comparação, determina-se a medida de similaridade mais eficaz e o valor ideal de $k$ para *playlists* similares que será empregado no Sistema de Recomendação.

**Exemplo 2.4.6.3** *O Fluxograma apresentado na Figura 6, apresenta o processo da técnica leave-one-out utilizado neste trabalho.*

No exemplo considerado, abordamos quatro *playlists* e seis músicas, aplicando a técnica do *Leave-One-Out* à Matriz de Utilidade original. Nesse contexto, as *playlists* representam os usuários, e as músicas são os itens, tendo como objetivo avaliar a eficácia das medidas de similaridade na previsão de entradas alteradas. Essa abordagem envolve a seleção arbitrária de entradas marcadas como 1, substituindo-as por 0 para examinar como diferentes medidas de similaridade, considerando variados valores de $k$ *playlists* vizinhas, sugerem músicas que originalmente foram consumidas.

O processo inicia-se ao escolher a *Playlist 1*, retirando duas entradas marcadas como 1 e transformando-as em 0. Em seguida, aplicamos diversas medidas de similaridade, variando $k$ *playlists* vizinhas, para obter estimativas. Repetimos esse procedimento para todas as *playlists*, gerando estimativas para cada medida de similaridade e diferentes valores de $k$. Posteriormente, consolidamos essas estimativas, apresentando-as visualmente

por meio de *boxplots* para facilitar a comparação entre as medidas de similaridade e seus respectivos $k$ *playlists* vizinhas.

Com a matriz de utilidade original, o primeiro passo consiste na escolha aleatória de uma entrada de consumo da *Playlist 1*, seguido pela alteração dessa entrada para 0 no segundo passo. Em seguida, no terceiro passo, calculamos diferentes similaridades para diversos valores de $k$. Repetimos esse processo para a *Playlist 2* no quarto passo e para todas as outras *playlists*, gerando resultados plotados em *boxplots*. Finalmente, no oitavo passo, analisamos minuciosamente os resultados para determinar a medida de similaridade que apresentou melhor desempenho em termos de precisão e relevância para o nosso Sistema de Recomendação. Este processo rigoroso assegura uma escolha criteriosa dos hiperparâmetros, alinhando o sistema ao contexto da Filtragem Colaborativa com dados binários.

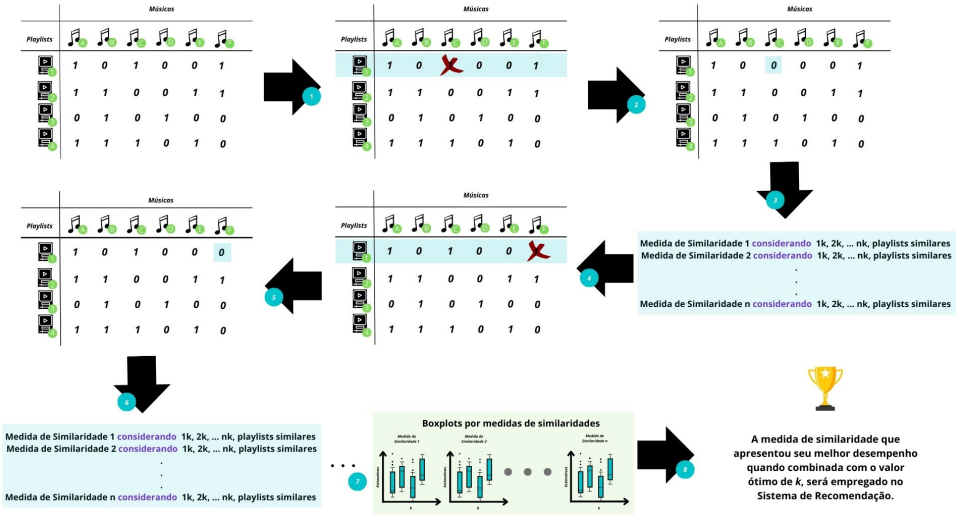


Figura 9: Fluxograma da técnica Leave-one-out

# 3 Análise dos Resultados

Neste capítulo estão apresentados as análises de resultados. As análises foram realizadas utilizando o software estatístico R (R Core Team, 2021).

## 3.1 Análise Descritiva

Os dados utilizados para este trabalho foram extraídos do *Million Playlist Dataset* (MPD), cedido pelo *Spotify* para a competição da *RecSys*[8] de 2018. O objetivo deste campeonato era realizar recomendações de músicas após o consumo completo de uma determinada *playlist*. Com isso, os consumidores passariam mais tempo utilizando a plataforma aumentando a popularidade e o consumo do *Spotify*. O banco de dados contém *playlists* de consumidores residentes dos Estados Unidos, com pelo menos 13 anos de idade e os dados foram coletados no período de 1º de Janeiro de 2010 a de 1º de Dezembro de 2017. As *playlists* eram públicas, com no mínimo 5 músicas e no máximo 250 músicas, com 3 artistas e 2 álbuns únicos.

Para a construção do Sistema de Recomendação, foi utilizada uma subbase com 66.721 músicas. Na Tabela 5, pode-se observar as estatísticas sobre as *playlists*.

Tabela 5: Estatísticas sobre as *playlists*

| Mínimo | 1º quartil | Mediana | Média | 3º quartil | Máximo |
|---|---|---|---|---|---|
| 5 | 28 | 52 | 76 | 104 | 1725 |

Após o tratamento para retirar playlists com nomes e músicas iguais a base passou a ter 869 *playlists* únicas e 34.443 músicas distintas.

[8]https://recsys.acm.org/recsys18/challenge/

## 3.2 Filtragem Colaborativa

Neste trabalho, será desenvolvido um Sistema de Recomendação de Filtragem Colaborativa em que os dados são representados de forma binária. O usuário é representado por *playlists* e o item é representado pelas músicas, sendo os dados binários utilizados para indicar a presença ou ausência de uma música em uma determinada playlist. Essa abordagem permite uma representação eficiente e compacta das interações entre usuários e itens, simplificando o processo de recomendação. O desafio está em como utilizar de forma eficaz esses dados binários na Filtragem Colaborativa, aprimorando a precisão das recomendações e proporcionando uma experiência musical enriquecedora aos consumidores do sistema.

A Matriz de Utilidade construída a partir da base de dados que possui dimensão de 869 $\times$ 34443, na qual cada vetor-linha representa uma *playlist* e cada vetor-coluna representa uma música. Na Tabela 6, por exemplo, estão representadas 5 *playlist* e 5 músicas da Matriz de Utilidade.

Tabela 6: Materiz de Utilidade: Playlists X Músicas

| *Playlist* | Música 1 | Música 2 | Música 3 | Música 4 | Música 5 |
|---|---|---|---|---|---|
| *Throwbacks* | 1 | 1 | 1 | 1 | 1 |
| *w o r k o u t* | 1 | 1 | 0 | 0 | 0 |
| *party playlist* | 1 | 1 | 0 | 0 | 0 |
| *Dance mix* | 1 | 0 | 0 | 0 | 0 |
| *spin* | 1 | 1 | 0 | 0 | 0 |

As entradas iguais a 1 representam que a música foi consumida pela *playlist* e a entrada 0 pode ser intepretada como uma música desconhecida pelo consumidor e por isso não está na *playlist* ou apenas porque o consumidor apesar de conhecer a musica optou em não ineseri-la. A medida de similaridade tem um importante papel em definir quais *playlist* são similares para calcular as previsões para a recomendação de músicas.

### 3.2.1 Comparação entre as Medidas de Similaridades

O Sistema de Recomendação via Filtragem Colaborativa é construído através da Matriz de Utilidade (Seção 2.4.3) e a Matriz de Similaridade entre Usuários ou Itens (Seção 2.4.4). Neste trabalho, foram estudadas diferentes abordagens de medidas de similaridades. Estas medidas são a Similaridade de *Jaccard*, *3W-Jaccard*, *Sorensen-Dice*, *Ochai*, S*okal & Sneath* e *Sokal & Michener*.

O objetivo desta pesquisa consiste em desenvolver um Sistema de Recomendação (SR) baseado em Filtragem Colaborativa, utilizando a Similaridade entre Usuários para recomendar músicas a *playlists* específicas. A construção do SR envolve a avaliação criteriosa da medida de similaridade mais adequada para os dados binários, visando garantir a precisão e a relevância das recomendações fornecidas pelo sistema. Além disso, é necessário determinar o número ideal de *playlists* similares a serem consideradas como referência durante o processo de recomendação, com o intuito de assegurar a diversidade e a qualidade das indicações musicais. A presente pesquisa propõe uma análise sistemática desses dois aspectos, com o objetivo de otimizar a eficácia do Sistema de Recomendação e fornecer recomendações musicais mais precisas e satisfatórias para as *playlists* de interesse. Para tanto, é importante selecionar uma medida de similaridade que seja capaz de expressar adequadamente a semelhança entre as playlists, levando em consideração as particularidades dos dados binários. Essas particularidades decorrem das entradas com valor 0, que podem indicar tanto a ausência de uma música na *playlist* quanto o desconhecimento da música pelo usuário. Portanto, para a seleção da medida de similaridade mais adequada e o número de *playlists* similares para a elaboração do Sistema de Recomendação, será conduzido um cenário de simulação no qual serão calculadas diferentes medidas para diferentes valores de "k"playlists mais próximas, conforme abordado na Seção 2.4.6.1.

A primeira etapa é que através do estimador *Leave-One-Out*(2.4.6.1), será encontrada as *playlists* com maior similaridade. Devido a limitações computacionais será selecionado apenas duas entradas aleatoriamente iguais a 1 em cada *playlist* da base de dados. Estas entradas serão modificadas para 0 com o intuito de futuramente verificar se as estimativas ficarão próximas de 1 indicando que a musica dada como não comsumida será recomendada como deveria. Após a construção das listas de *playlists* mais semelhantes será fixado $k$ de diferentes valores para caclular a estimativa baseada nas *playlists* mais próximas. Os valores escolhidos são $\mathbf{k = 5}$,$\mathbf{k = 10}$, $\mathbf{k = 20}$ $\mathbf{k = 50}$, $\mathbf{k = 100}$, $\mathbf{k = 500}$, $\mathbf{k = 868}$. A partir dos valores de $\mathbf{k}$ definidos será calculado a estimativa, conforme a Seção **??**, para obter a previsão se a música foi recomendada como o esperado.

A Figura 10 apresenta os resultados das Similaridades de *Jaccard*, *3W-Jaccard*, *Sorensen-Dice*, *Ochai*, *Sokal & Sneath* e *Sokal & Michener*.

Ao observar as medidas das Similaridades de *Jaccard* e *3W-Jaccard*, pode-se observar que ambas mostraram o mesmo comportamento em relação aos valores de $\mathbf{k}$. Vale ressaltar que o *boxplot* referente ao $\mathbf{k = 5}$ apresentou desempenho superior aos demais e por outro lado o *boxplot* de $\mathbf{k = 868}$ que considerou toda a base teve estimativa menores.

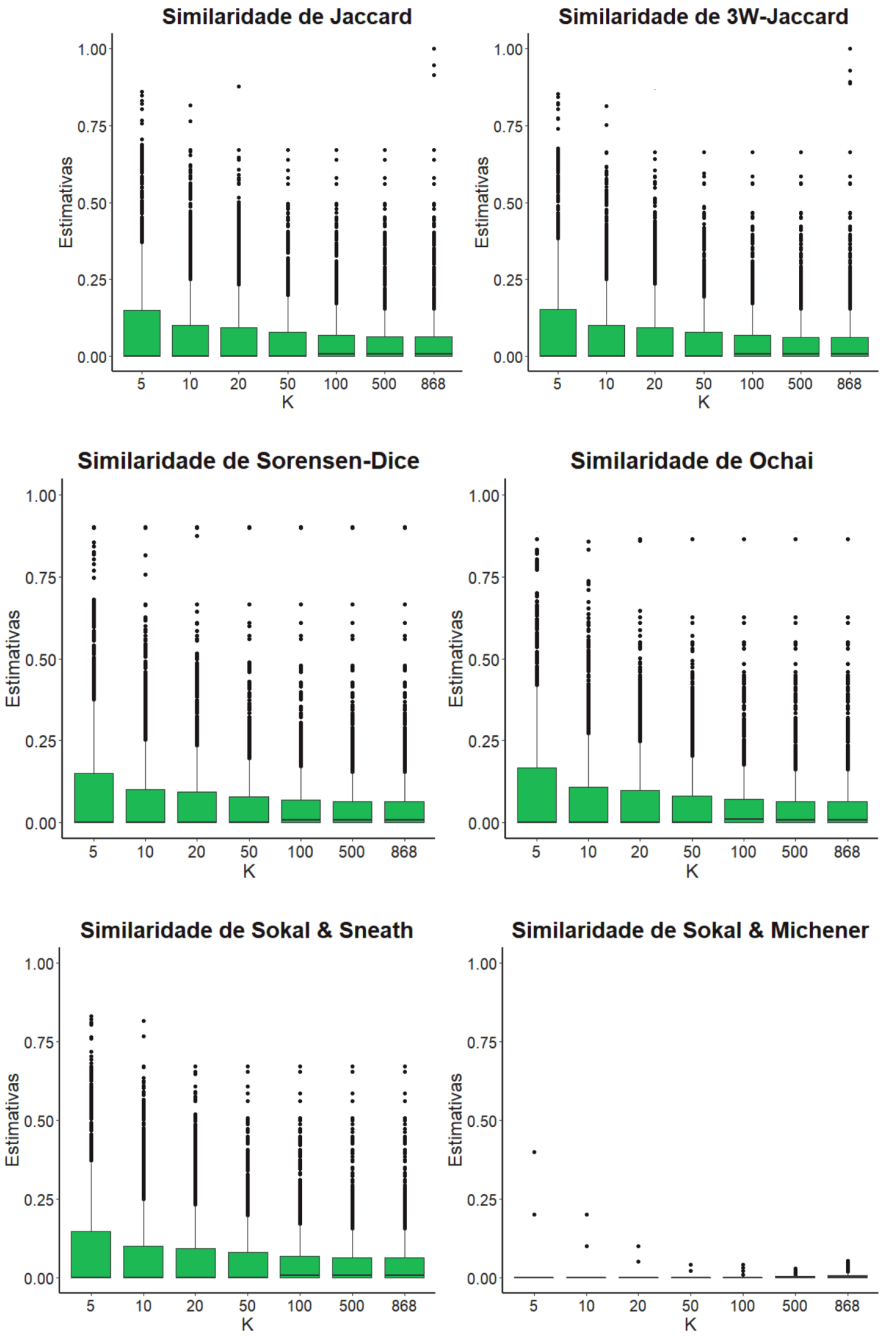


Figura 10: Estimativas para Similaridades de *Jaccard*, *3W-Jaccard*, *Sorensen-Dice*, *Ochai*, *Sokal & Sneath* e *Sokal & Michener*

Assim como na anterior pode-se notar que as medidas de Similaridades de *Sorensen-Dice* e *Ochai* também apresentaram comportamentos bem parecidos apesar dos *outlines* serem um pouco diferente. Além disso, também pode-se destacar que o valor de $\mathbf{k = 5}$ também foi o com melhor desempenho.

Por fim, foi observado que a Similaridade de *Sokal & Sneath* apresentou valores para $\mathbf{k = 5}$ próximos das medidas de similaridade discutidas anteriormente, conforme eviden-

ciado pelos *boxplots*. Já a Similaridade de *Sokal & Michener*, foi a única das medidas que obtive as estimativas bem próximas de 0 indicando que a medida não expressa bem a similaridade entre as playlists afetando o resultado final das estimativas, como foi visto na Seção (2.4.5.5).

Com o cenário de simulação finalizado, será definida a medida de similaridade e o valor **k** que tiveram melhor desempenho. Das medidas analisadas a Simialridade de *Sokal & Michener* foi aquela com pior desemenho nos valores das estimativas calculadas. Isto era esperado pois esta medida considera as ausências conjuntas, ou seja, irá considerar as músicas que não estão presentes em ambas as *playlists*.

O valor de $\mathbf{k = 5}$ foi o que teve melhor desempenhuo dentro os **k** avaliados. Portanto, a Similaridade *3W-Jaccard* com $\mathbf{k = 5}$ é mais adquado para o Sistema de Recomendação que está sendo proposto.

As medidas de similaridade apresentaram nos seus receptivos *boxplots* valores próximos. A escolha da Similaridade de *3W-Jaccard* para trabalho é justificada pela sua resistência a dados esparsos e invariância à escala, são atributos cruciais para a eficácia no tratamento de conjuntos tão numerosos, como aqueles encontrados nas *playlists* do *Spotify*. Comparativamente, outras métricas, como *Jaccard*, *Sorensen-Dice*, *Ochai*, *Sokal Sneath*, e *Sokal Michener*, podem apresentar complexidades desnecessárias ou não oferecer vantagens substanciais que justifiquem sua escolha sobre *3W-Jaccard*. A eficiência computacional da Similaridade de *3W-Jaccard* é particularmente relevante, considerando a aplicação em sistemas de recomendação que frequentemente lidam com grandes conjuntos de dados como estamos usando no trabalho. Além disso, foi definido o $k$ ótimo como 5 porque apresentou o melhor desempenho dentro os avaliados com a melhor amplitude.

A Matriz de Similaridade entre Usuários será calculada com a Similaridade de *3W-Jaccard*) com dimensão $869 \times 869$. A medida escolhida foi a Similaridade de *3W-Jaccard* para construir o Sistema de Recomendação. Na Tabela 7, está representada a similaridade entre as *playlists*(usuários).

Tabela 7: As 5 primeiras linhas e colunas da Materiz de Similaridade

| | *Throwbacks* | *Awesome Playlist* | *korean* | *mat* | *90s* |
|---|---|---|---|---|---|
| *Throwbacks* | 1 | | | | |
| *Awesome Playlist* | 0.000000 | 1 | | | |
| *korean* | 0.000000 | 0.000000 | 1 | | |
| *mat* | 0.000000 | 0.000000 | 0.000000 | 1 | |
| *90s* | 0.058632 | 0.000000 | 0.000000 | 0.000000 | 1 |

A Matriz de Similaridade entre Usuários (Tabela 7) será utilizada nas recomendações de músicas para *playlists* da Seção 3.2.2.

### 3.2.2 Recomendação a partir da Similaridade de *3W-Jaccard* entre Usuários

A recomendação das músicas serão feitas para três *playlists* utilizando a Similaridade de *3W-Jaccard* entre Usuários e $\mathbf{k = 5}$. Para determinar o número de músicas a serem recomendadas em cada *playlist*, foi utilizada a mesma ideia do *Spotify*, em que recomenda 5 músicas semelhantes conforme a *playlist* consumida. Para analisar descritivamente os resultados foram selecionada *playlists* com um determinado gênero para verificar se o Sistema de Recomendação fará recomendações de músicas com o mesmo gênero.

A primeira *playlist* a receber recomendações de músicas será a nomeada como *Rap*, que contém 32 músicas do gênero *rap*. A similaridade de *3W-Jaccard* calculará a similaridade entre a *playlist Rap* com as demais *playlists* da base de dados. A Tabela 8, apresenta as 5 *playlists* mais semelhantes a *playlist Rap*:

Tabela 8: As 5 *playlists* mais semelhantes em relação a *playlist Rap*

| *Playlists* | Similaridade de *3W-Jaccard* |
|---|---|
| *Chill Out* | 0.130435 |
| *Sad songs* | 0.095238 |
| *My Heart* | 0.084507 |
| *Summer drives* | 0.081818 |
| *Yeet* | 0.078261 |

Com o método de vizinhos próximos, foram listadas as 5 *playlists* mais próximas da *playlist Rap*. Com base nelas será feita o cálculo das estimativas para obter as músicas a serem recomendadas. Na Tabela 9, será selecionada as 5 músicas com maiores estimativas.

Tabela 9: As 5 músicas recomendas para *playlist Rap*

| Músicas | Artista | Chance de Consumo |
|---|---|---|
| *Luxury* | Jon Bellion | 0.2773680 |
| *No Role Modelz* | J. Cole | 0.2773680 |
| *IV. sweatpants* | Childish Gambino | 0.2773680 |
| *Bonfire* | Childish Gambino | 0.2773680 |
| *V. 3005* | Childish Gambino | 0.2773680 |

As músicas *Luxury*, *No Role Modelz*, *IV. sweatpants*, *Bonfire* e *V. 3005* foram recomendadas a *playlist Rap*. É importante notar que as músicas recomendadas são do gênero *rap* e o cantor *Childish Gambino* que aparece 3 vezes na recomendação é um cantor presente na *playlist Rap* com outras músicas. Logo, pode-se perceber que as músicas recomendadas tem grandes chances de satisfazer o consumidor da *playlist Rap*.

A próxima *playlist* é a *Christmas Favorites* com 64 músicas com temas natalinos. As 5 *playlists* mais próximas e o valor das similaridades estão na Tabela 10:

Tabela 10: As 5 *playlists* mais semelhantes em relação a *playlist Christmas Favorites*

| *Playlists* | Similaridade de *3W-Jaccard* |
|---|---|
| *CHRISTMAS* | 0.411290 |
| *Xmas* | 0.336585 |
| *Christmas* | 0.314286 |
| *CHRISTMAS MUSIC* | 0.3 |
| *Christmas 2016* | 0.25 |

Pode-se observar que as 5 *playlists* mais similares da *playlist Christmas Favorites* também são *playlists* com músicas natalinas.

Tabela 11: As 5 músicas recomendas para *playlist Christmas Favorite*

| Músicas | Artista | Chance de Consumo |
|---|---|---|
| *Christmas Wrapping* | The Waitresses | 0.2551173 |
| *I Saw Mommy Kissing Santa Claus* | The Jackson 5 | 0.2551173 |
| *Rudolph the Red-Nosed Reindeer* | Gene Autry | 0.2551173 |
| *Have Yourself A Merry Little Christmas* | Michael Bublé | 0.2551173 |
| *Here Comes Santa Claus (Right Down Santa Claus Lane) - Single Version* | Bing Crosby | 0.2551173 |

As músicas *Christmas Wrapping*, *I Saw Mommy Kissing Santa Claus*, *Rudolph the Red-Nosed Reindeer*, *Have Yourself A Merry Little Christmas* e *Here Comes Santa Claus (Right Down Santa Claus Lane) - Single Version & Bing Crosby* foram recomendadas a *playlist Christmas Favorite*. Todas as músicas recomendadas são músicas que remetem as Festas de Natal.

Por fim, a *playlist Classic* com 204 músicas clássicas de *rock*. Esta *playlist* precisa receber recomendações de músicas do gênero de *rock* e ela também precisa ser um clássico. As *playlists* mais similares será apresentada na Tabela 12:

Tabela 12: As 5 *playlists* mais semelhantes em relação a *playlist Classic*

| *Playlists* | Similaridade de *3W-Jaccard* |
|---|---|
| *PARTY* | 0.356688 |
| *classic* | 0.356061 |
| *Classic Rock* | 0.289256 |
| *Good Stuff* | 0.286957 |
| *Elisa* | 0.279793 |

Na Tabela 13, as 5 músicas recomendadas a *playlist Classic* foram *Paradise City*, *Sweet Home Alabama*, *Knockin' On Heaven's Door*, *Dream On* e *Runnin' With The Devil - 2015 Remastered Version* que são do gênero *rock* e também são músicas classicas do gênero.

Tabela 13: As 5 músicas recomendas para *playlist Classic*

| Músicas | Artista | Chance de Consumo |
|---|---|---|
| *Paradise City* | Guns N' Roses | 0.817080 |
| *Sweet Home Alabama* | Lynyrd Skynyrd | 0.817080 |
| *Knockin' On Heaven's Door* | Guns N' Roses | 0.638727 |
| *Dream On* | Aerosmith | 0.638727 |
| *Runnin' With The Devil - 2015 Remastered Version* | Van Halen | 0.638727 |

Analisando as músicas recomendadas paras três *playlist* determinadas existe a possibilidade do Sistema de Recomendação baseado em Filtragem Colaborativa utilizando a medida de *3W-Jaccard* seja adequada para os dados binários da estudados neste trabalho.

# 4 Conclusão

O presente trabalho desenvolveu um Sistema de Recomendação por Filtragem Colaborativa através de uma base de dados binárias refente ao consumo de músicas por *playlists*. Dentre as metodologias utilizadas dentro do sistema pode-se destacar a Similaridade entre Usuários com as seguintes medidas de similaridades *Jaccard* e *3W-Jaccard*, *Sorensen-Dice*, *Ochai* , *Sokal & Sneath* e *Sokal & Michener* e o uso de difentes **k** para determinar os usuários mais próximos. A base de dados utilizada contém 1000 *playlists* e 66721 músicas, que é uma das bases disponibilizadas pelo *Spotify* na *RecSys* de 2018. Os consumos são definidos das seguintes maneiras: 0 - se o item não foi consumido por desconhecimento de sua existência pelo o consumidor ou ausência em interesse em consumi-lo, 1 - se o item foi consumido.

Para que seja feita a Filtragem Colaborativa, foi necessário tratar a base de dados desconsiderando as *palylists* com mesmo nome e mesmas músicas. Assim a base de dados passou a ter 869 *palylists* distintas e 34443 músicas únicas. Desta forma, foi utilizada esta base de dados para a elaboração do Sistema de Recomendação.

A fim de comparar as medidas de similaridades, foi realizado um cenário de simulação para definir a medida e **k** que apresentam melhores desempenhos. Primeiramente, foram selecionadas ao acaso duas músicas consumidas em cada *playlist* da base de dados e em seguidas estas entradas foram modificadas para 0, ou seja, foram consideradas como informação desconhecida. Utilizando o método de *Leave-one-out* foi feito o cálculo de Similaridade entre Usuários para definir as *playlists* mais similares. Para calcular as estimativas foi preciso definir que os **k** vizinhos mais próximos seriam $\mathbf{k = 5}$, $\mathbf{k = 10}$, $\mathbf{k = 20}$, $\mathbf{k = 50}$, $\mathbf{k = 100}$, $\mathbf{k = 500}$, $\mathbf{k = 868}$ para prever os valores para estas entradas alteradas. A Similaridade de *3W-Jaccard* foi com melhor desempenho devido aos pesos maiores que são dadas as correspondências conjuntas e o $\mathbf{k = 5}$ mostrou que ao considerar somente as *playlist* bem mais próximas o a previsão fica mais próxima de 1.

Após o processo de definir a medida de similaridade *3W-Jaccard* e o **k** vizinho mais

próximo, para realizar a recomendação foram selecionadas 3 *playlists* para receberem a 5 músicas semelhantes a músicas da *playlist* consumida. A primeira *playlist* escolhida foi a *rap* com 32 músicas e que obteve as seguintes *playlists* mais próximas com o respectivo valor da similaridade: *Chill Out* = 0.130435, *Sad songs* = 0.095238, *My Heart* = 0.081818, *Summer drives* = 0.081818 e *Yeet* = 0.078261. As músicas recomendas para *rap* foram *Luxury*, *No Role Modelz*, *IV. sweatpants*, *Bonfire* e *V. 3005*, que são do gênero *rap* como o esperado. A segunda *playlist* foi a *Christmas Favorites* com 64 músicas natalinas e que obteve as seguintes *playlists* mais próximas com o respectivo valor da similaridade: *CHRISTMAS* = 0.411290, *Xmas* = 0.336585, *Christmas* = 0.314286, *CHRISTMAS MUSIC* = 0.3 e *Christmas 2016* = 0.25. As músicas recomendas para *Christmas Favorites* foram *Christmas Wrapping*, *I Saw Mommy Kissing Santa Claus*, *Rudolph the Red-Nosed Reindeer*, *Have Yourself A Merry Little Christmas* e *Here Comes Santa Claus (Right Down Santa Claus Lane) - Single Version & Bing Crosby*, que são todas natalinas como na *Christmas Favorites*. A terceira *playlist* a receber recomendações foi *Classic* com 204 músicas do gênero *rock* antigo e que obteve as seguintes *playlists* mais próximas com o respectivo valor da similaridade: *PARTY* = 0.356688, *classi* = 0.356061, *Classic Rock* = 0.289256, *Good Stuff* = 0.286957 e *Elisa* = 00.279793. As 5 músicas recomendadas a *Classic* foram *Paradise City*, *Sweet Home Alabama*, *Knockin' On Heaven's Door*, *Dream On* e *Runnin' With The Devil - 2015 Remastered Version* que são do gênero *rock* antigo.

Com base nas recomendações realizadas para três *playlists*, pode-se observar um bom desempenho do Sistema de Recomendação elaborado considerando que recomendou as músicas com gêneros semalhantes ou iguais. Uma forma de avaliar o sistema seria de fato com um consumidor indicando que gostou ou não das músicas recomendadas.

Conclui-se que ao construir um Sistema de Recomendação com dados binários é necessário levar em conta as peculiaridades das entradas 0 que tem duas formas de serem intepretadas (0 - não consumiu porque desconhece o item ou não consumiu por falta de interesse). Além disso, escolher a melhor medida de similaridade e o valor de k vizinhos mais próximos para obter bons resultados. Em trabalhos futuros, pode-se explorar outras medidas de similaridades voltadas para dados binários, novos valores para o k vizinhos mais próximos e principalmente expandir a recomendação para outros cenários.

# Referências

ALBUQUERQUE, M. A. et al. Comparação entre coeficientes similaridade um aplicação em ciências florestais. *Matemática e Estatística em Foco*, v. 4, n. 2, p. 102–114, 2016.

AZAMBUJA, R. X. d.; MORAIS, A. J.; FILIPE, V. Teoria e prática em sistemas de recomendação. *Revista de Ciências da Computação*, Universidade Aberta, p. 23–46, 2021.

BALADY, V. Novas perspectivas para a publicidade na era dos streamings. um estudo de caso sobre o spotify. *Educação, Cultura e Comunicação*, v. 11, n. 22, 2020.

FORMIGA, D. d. A. Suggestme: um sistema de recomendação utilizando web semântica para evitar o cold start. Universidade Federal da Paraíba, 2014.

GODINHO, A.; VASCONCELOS, F. Recomendação musical para grupos baseada em modelo híbrido.

GÓIS, M. d. M. Melhorias para um sistema de recomendação baseado em conhecimento a partir da representação semântica de conteúdos. Universidade do Vale do Rio dos Sinos, 2015.

GOMES, A. d. A. et al. Uma análise da qualidade de recomendações via filtragem colaborativa e regressão logística. 2019.

LIKERT, R. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

LIMA, T. A. S. Análise de risco em recomendaçoes de filmes e piadas via filtragem colaborativa. 2019.

MASSON, M. M. *Cold Start em Recomendação de Músicas Utilizando Deep Learning*. Tese (Doutorado) — PUC-Rio, 2016.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, Manole, v. 1, n. 1, p. 32, 2003.

NAPOLITANO, M. História &música–história cultural da música popular–belo horizonte. autêntica, 2002. _. *Cultura brasileira: utopia e massificação (1950/1980)*, v. 4, 2015.

RAMÍREZ, I.; BATYRSHIN, I. Análisis de propiedades de medidas de similitud con atributos binarios. *Research in Computing Science*, v. 116, p. 117–124, 2016.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to recommender systems handbook. In: *Recommender systems handbook*. [S.l.]: Springer, 2011. p. 1–35.

ROHLF, F. J.; SOKAL, R. R. The description of taxonomic relationships by factor analysis. *Systematic Zoology*, JSTOR, v. 11, n. 1, p. 1–16, 1962.

ROZENDO, R. G. *Avaliação de sistemas de recomendação com uma proposta de um algoritmo híbrido*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2017.

SILVA, J. F. G. da. *Efeitos da Esparsidade em Sistemas de Recomendaçao*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2019.

SOUSA, P. S. de; FERREIRA, A. A. Estimating similarity among entities aided by the web when only the entity name is available. In: *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*. [S.l.: s.n.], 2018. p. 253–260.

SOUZA, B. F. M.; MAILIDÚ, R. L. *Modelos de Fatoração Matricial para Recomendação de Vídeos. 2011. 67f.* Tese (Doutorado) — Dissertação-Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2011.

VERMA, V.; AGGARWAL, R. K. A comparative analysis of similarity measures akin to the jaccard index in collaborative recommendations: empirical and theoretical perspective. *Social Network Analysis and Mining*, Springer, v. 10, n. 1, p. 1–16, 2020.

WOLF, C. F. *Aquatic macroinvertebrate diversity and water quality of urban lakes*. Tese (Doutorado) — Texas Tech University, 1996.