UNIVERSIDADE PRESBITERIANA MACKENZIE

FACULDADE DE COMPUTAÇÃO E INFORMÁTICA

Sistema de Recomendação de Jogos na Plataforma Steam

Nome: Camila Vieira
RA: 10414794
E-mail: 10414794@mackenzista.com.br

Nome: Gabriel Schonenberger de Campos
RA: 10415150
E-mail: 10415150@mackenzista.com.br

Nome: Glayton de Paula
RA: 10415099
E-mail: 10415099@mackenzista.com.br

Nome: João Victor Mendes Cunha
RA: 10415459
E-mail: 10415459@mackenzista.com.br

São Paulo
2024

Sumário

1. INTRODUÇÃO:

Sistema de Recomendação de Jogos na Plataforma Steam

1.1 RESUMO

Com o avanço da indústria de jogos digitais, a plataforma Steam se destaca como uma das maiores distribuidoras de jogos para PC, oferecendo uma vasta biblioteca e uma rica base de dados sobre o comportamento dos usuários. Este trabalho propõe o desenvolvimento de um sistema de recomendação de jogos utilizando dados implícitos, como tempo de jogo e comportamento de acesso, com o objetivo de melhorar a experiência do usuário e aumentar o engajamento na plataforma.

A metodologia adotada baseia-se na utilização de matriz de correlação para identificar associações entre os jogos, empregando normalização de dados para lidar com a variabilidade e exclusão de usuários com interações mínimas. O modelo foi avaliado em termos de acurácia, atingindo 10% de sucesso na recomendação de jogos adquiridos, o que aponta para limitações em cenários mais complexos.

Os resultados destacam o potencial da abordagem para contextos simples, mas sugerem a necessidade de explorar métodos mais avançados, como filtragem colaborativa e modelos híbridos, para melhorar a personalização e a relevância das recomendações. Este trabalho contribui para o aprimoramento de sistemas de recomendação na indústria de jogos, oferecendo insights sobre os desafios e possibilidades dessa tecnologia.

1.2 CONTEXTO DO TRABALHO:

Com o crescimento da indústria de jogos digitais, a plataforma Steam se destaca como uma das maiores distribuidoras de jogos para PC, oferecendo uma vasta biblioteca que abrange diversos gêneros e estilos. A plataforma não apenas comercializa jogos, mas também armazena uma rica base de dados sobre o comportamento dos usuários, incluindo acessos, jogos jogados, tempo de jogo e avaliações.

Esses dados são fundamentais para desenvolver sistemas de recomendação que possam guiar os usuários na escolha de novos jogos, maximizando a experiência de jogo e aumentando o engajamento na plataforma. As recomendações baseadas em

comportamento permitem que os usuários descubram novos conteúdos alinhados aos seus interesses, tornando-se um diferencial competitivo para a Steam.

1.3 MOTIVAÇÃO:

A motivação para este trabalho reside na necessidade de melhorar a experiência dos usuários da Steam através de recomendações mais precisas e personalizadas. Um sistema de recomendação eficiente pode aumentar significativamente a satisfação do usuário, auxiliando na descoberta de novos jogos, e, consequentemente, impulsionando as vendas e o tempo de uso na plataforma.

Além disso a pesquisa se relaciona com os temas extensionistas de incentivo à cultural e ao bem-estar tendo em vista que cada vez mais o mercado de jogos está presente na vida das pessoas como uma fonte de entretenimento e cultura.

1.4 JUSTIFICATIVA

Os sistemas de recomendação são amplamente utilizados em diversas plataformas digitais, como e-commerce, serviços de streaming e redes sociais, para sugerir conteúdos relevantes aos usuários. Na Steam, essas recomendações têm o potencial de enriquecer a experiência de jogo ao sugerir títulos que o usuário possivelmente desconhece, mas que são altamente relevantes para o seu perfil de comportamento.

Este projeto busca explorar a aplicação de técnicas de aprendizado de máquina e análise de dados para a criação de um sistema de recomendação que vá além dos métodos tradicionais, utilizando não apenas as avaliações e compras, mas também o comportamento de acesso e tempo de jogo dos usuários.

1.5 OBJETIVOS E METAS

Objetivo Principal: Desenvolver um sistema de recomendação de jogos para a plataforma Steam utilizando dados comportamentais dos usuários, como jogos jogados, tempo de jogo e acessos.

Metas:

- Coletar e processar dados de comportamento dos usuários da Steam.
- Analisar padrões de comportamento de jogadores para identificar preferências e tendências.
- Desenvolver algoritmos de recomendação que utilizem aprendizado de máquina para sugerir jogos.
- Validar o modelo de recomendação com base na precisão e relevância das sugestões feitas aos usuários.

2. REFERENCIAL TEÓRICO

De forma geral, os sistemas de recomendação são baseados em duas estratégias diferentes (ou combinações delas). A abordagem **baseada em conteúdo (contente based)** cria um perfil para cada usuário ou produto para caracterizar sua natureza. Por exemplo, o perfil de um filme pode incluir atributos relacionados ao seu gênero, os atores participantes, sua popularidade nas bilheterias etc. Os perfis dos usuários podem incluir informações demográficas ou respostas a um questionário adequado. Os perfis resultantes permitem que os programas associem usuários a produtos correspondentes. Esse tipo de dado é chamado de **dado explícito**, quando o usuário tem informações que permitem linkar este a um produto que também tem informações listadas e que se igualam às preferências do usuário. No entanto, as estratégias baseadas em conteúdo exigem a coleta de informações externas que podem não estar disponíveis ou ser difíceis de obter.

Uma estratégia alternativa, depende apenas do comportamento passado dos usuários, sem exigir a criação de perfis explícitos. Essa abordagem é conhecida como **Filtragem Colaborativa (Collaborative Filtering - CF)**, um termo cunhado pelos desenvolvedores do primeiro sistema de recomendação. A CF analisa as relações entre os usuários e as interdependências entre os produtos, a fim de identificar novas associações entre usuários e itens. Por exemplo, alguns sistemas de CF identificam pares de itens que tendem a ser avaliados de forma semelhante, ou usuários com opiniões semelhantes, com um histórico de avaliações ou compras semelhantes, para deduzir relações desconhecidas entre usuários e itens. A única informação necessária é o comportamento passado dos usuários, que pode incluir suas transações anteriores ou a forma como avaliam produtos. Esses dados são chamados de **dados implícitos**, pois são coletados de forma não declarada ao usuário. Um dos principais atrativos da CF é que ela é independente do domínio, mas pode abordar aspectos dos dados que muitas vezes são difíceis de conseguir usando técnicas baseadas em conteúdo. Embora, em geral, seja mais precisa do que as técnicas baseadas em conteúdo, a CF sofre com o problema do "cold start", devido à sua incapacidade de lidar com produtos novos no sistema, para os quais as abordagens baseadas em conteúdo seriam adequadas. Isso acontece, pois, um usuário novo, que ainda não interagiu com muitos produtos, não vai ter um perfil construído de recomendações, então suas recomendações inicialmente podem ser vagas.

Existem diversos modelos que podem ser utilizados para Filtragem Colaborativa (CF) com dados implícitos, cada um com suas vantagens e desvantagens. Um dos métodos mais comuns é a Fatoração de Matrizes, como o Alternating Least Squares (ALS), que é eficaz na identificação de padrões de comportamento dos usuários, permitindo uma boa escalabilidade, mas pode ser sensível a dados esparsos. Outro modelo popular é o Singular Value Decomposition (SVD) adaptado para dados implícitos, que oferece uma representação compacta das interações, mas pode ser mais complexo em termos de implementação e menos interpretável. O modelo de K-vizinhos mais próximos (KNN) é simples e fácil de entender, funcionando bem em datasets menores, mas pode ser menos eficiente em grandes escalas devido ao tempo de computação. Além disso, abordagens baseadas em aprendizado de máquina, como

redes neurais e matrizes de correlação, têm se mostrado promissoras por sua capacidade de capturar relações não lineares, mas requerem um volume considerável de dados para treinamento e podem ser difíceis de ajustar. Cada um desses métodos apresenta trade-offs que devem ser considerados com base nas características dos dados e nos objetivos específicos do sistema de recomendação.

A escolha do modelo vai interferir na qualidade final dos resultados e é comum empresas usarem mais de um método para seus sistemas de recomendação, potencializando seus resultados.

Sobre a avaliação dos resultados dos modelos de sistemas de recomendação é uma tarefa complexa que envolve diversas métricas e considerações. Uma das principais dificuldades é a natureza dos dados, especialmente quando se trabalha com dados implícitos, onde as interações não são sempre explícitas e os feedbacks dos usuários podem ser limitados. Além disso, as métricas comuns de avaliação, como precisão, recall e F1-score, precisam ser adaptadas para refletir a realidade dos dados, uma vez que a ausência de uma interação não significa necessariamente que o usuário não gostou do item.

Outro desafio é a escolha do conjunto de validação, que deve ser representativo e balanceado para evitar viés nos resultados. A divisão dos dados em conjuntos de treinamento e teste pode influenciar significativamente a avaliação, especialmente em cenários onde os dados são esparsos. Adicionalmente, a interpretação dos resultados pode ser complicada, uma vez que um modelo pode apresentar boa performance em termos de métricas quantitativas, mas falhar em fornecer recomendações que sejam relevantes ou satisfatórias para os usuários.

2.1 MÉTODO ESCOLHIDO: MATRIZ DE CORRELAÇÃO

A metologia escolhida para desenvolvimento do projeto foi a matriz de correlação. A matriz de correlação é uma ferramenta estatística amplamente empregada para medir a associação entre variáveis, indicando o grau de similaridade ou interdependência entre elas. Para isso, utiliza-se coeficientes de correlação, e aqui o foco é no coeficiente de Pearson. Esse coeficiente, também conhecido como “correlação produto-momento” ou “ρ de Pearson,” varia de -1 a 1, indicando a direção e a força da

associação entre as variáveis. Quanto mais próxima de 1, a correlação é mais forte, oque quer dizer que normalmente aquelas variáveis têm uma relação diretamente proporcional. Por outro lado, uma correlação próxima de -1 indica uma relação inversamente proporcional e, por fim, uma correlação negativa indica ausência de correlação entre as variáveis

Esse método oferece uma base sólida para recomendações, pois permite a exploração das associações diretas entre itens, sem exigir dados profundos de comportamento do usuário, como histórico extenso ou perfis detalhados. A simplicidade torna a matriz de correlação uma abordagem prática, especialmente em cenários onde o foco está na recomendação de itens populares ou na identificação de combinações recorrentes em um catálogo de produtos.

No contexto de um sistema de recomendação, a matriz de correlação é calculada com base nas interações ou características compartilhadas entre itens, possibilitando a identificação de padrões que sugerem quais itens são frequentemente consumidos em conjunto. No caso do modelo de recomendação construído, os dados utilizados foram dados implícitos, ou seja, queremos aqui criar recomendações com base no comportamento de usuários e não em dados catalogados.

3. METODOLOGIA

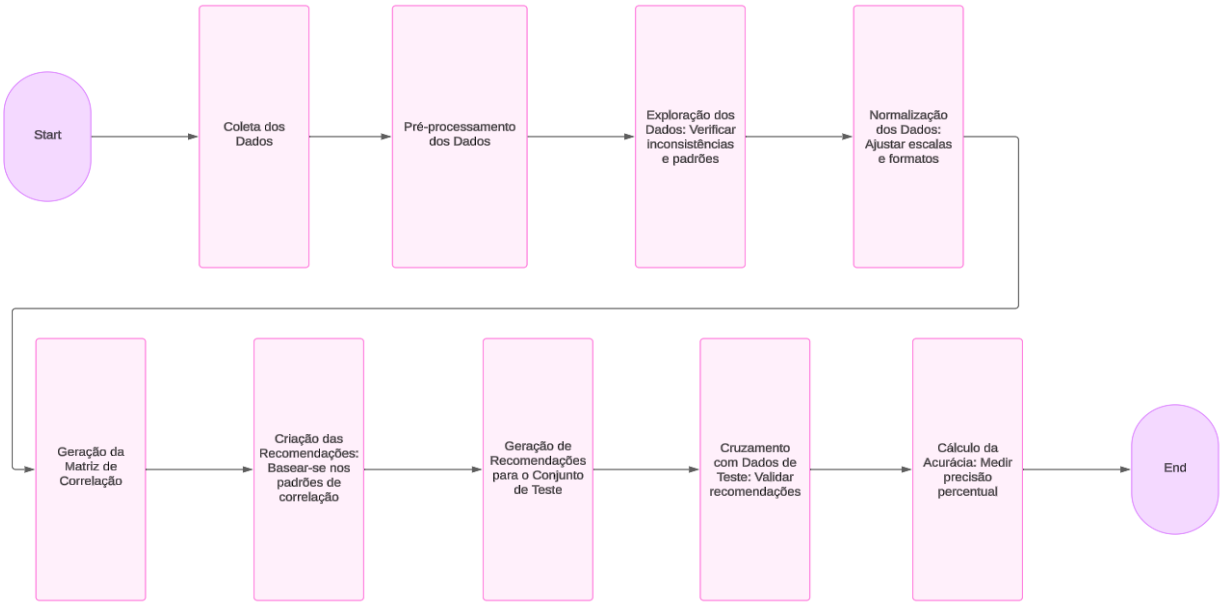
O desenvolvimento do sistema de recomendação seguiu uma abordagem estruturada, começando pela coleta dos dados. Para isso, utilizou-se uma base pública contendo informações sobre usuários, jogos e tempo de jogo, acessada por meio do repositório disponível no site da UCSD. Após a coleta, iniciou-se o processo de pré-processamento dos dados, etapa crucial para garantir a qualidade e a utilidade da base.

No pré-processamento, inicialmente foram exploradas as características gerais da base, como a distribuição do campo playtime_forever, que apresenta o tempo de jogo dos usuários. A análise revelou uma alta concentração de valores em torno de zero, evidenciando que muitos usuários jogaram pouco ou nenhum dos jogos presentes em suas contas. Para lidar com essa discrepância, aplicou-se uma normalização utilizando a transformação logarítmica (np.log10) para equilibrar a escala dos dados.

Além disso, foram filtrados usuários com menos de dois jogos registrados, pois não contribuíam de forma significativa para as análises de correlação e poderiam introduzir ruídos no modelo. Por fim, os dados foram divididos em conjuntos de treino e teste na proporção 80/20, assegurando que os mesmos usuários estivessem presentes em ambos os conjuntos para permitir a validação consistente do modelo.

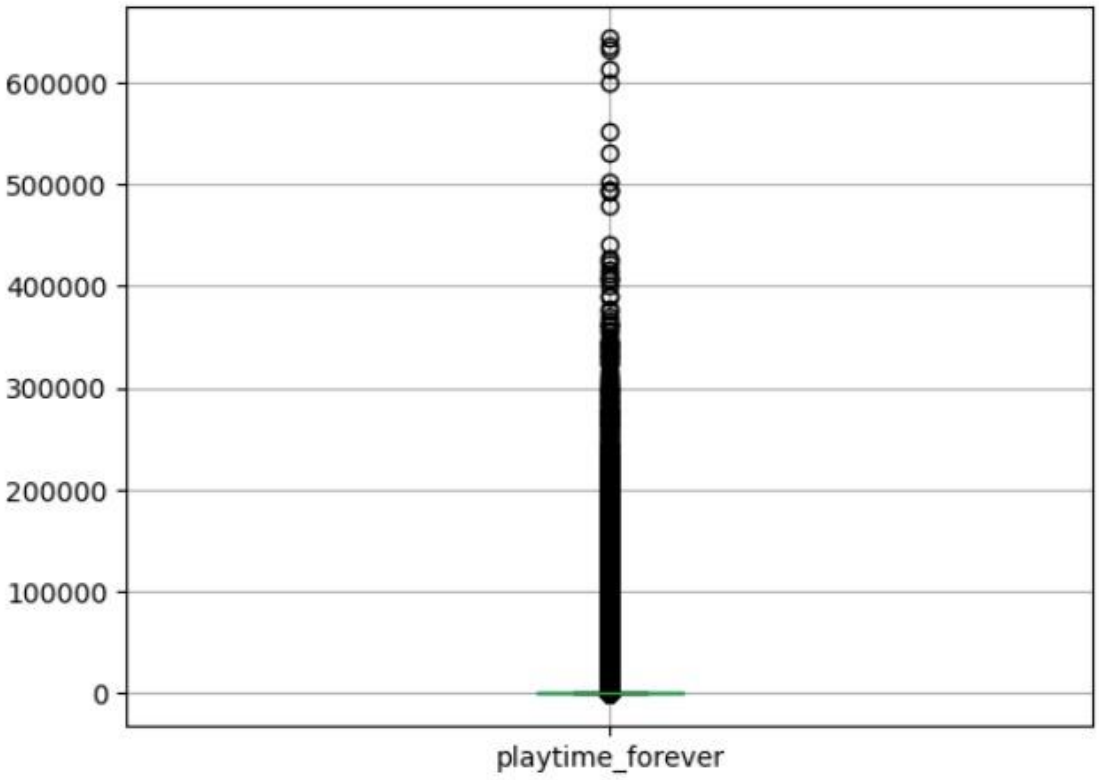
Com a base preparada, o próximo passo foi a construção do modelo de recomendação. A abordagem escolhida baseou-se na matriz de correlação, utilizando o coeficiente de Pearson para identificar associações entre os jogos. Essa matriz foi gerada a partir do conjunto de treinamento e apresentou dimensões expressivas (aproximadamente 10.000 x 10.000). Com a matriz pronta, desenvolveu-se um algoritmo para recomendar os 10 jogos mais relevantes para cada usuário. Para promover diversidade, consideraram-se os cinco jogos mais jogados por cada usuário e selecionaram-se os dois jogos com maior correlação para cada um deles.

A etapa final consistiu na avaliação do modelo. Foram geradas recomendações para todos os usuários do conjunto de teste e comparadas com os jogos efetivamente registrados como jogados. A figura abaixo mostra um diagrama geral da metodologia

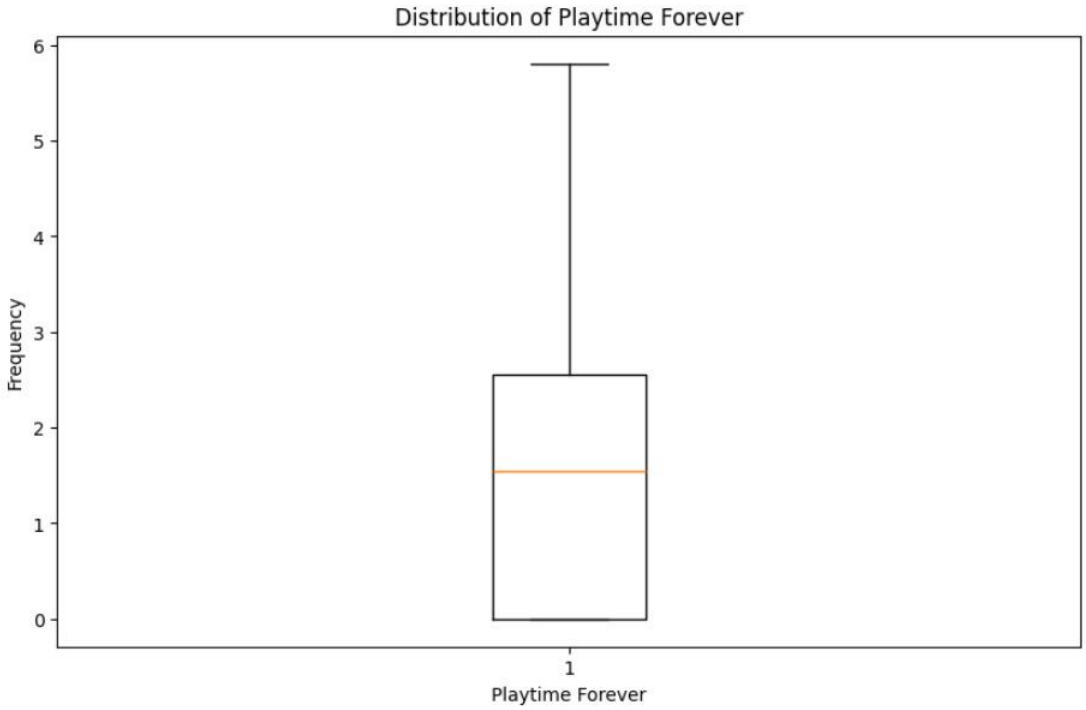


4. RESULTADOS

Em uma análise de correlação é muito importante avaliar a distribuição dos dados, para isso, foi gerado um boxplot do campo playtime_forever, que é o campo que efetivamente vai medir a correlação entre items. O gráfico ficou com a seguinte formação:



Ao observar a imagem, podemos perceber que a base de dados tem grande contraste entre os usuários, a maioria deles jogou 0 horas ou poucas horas dos jogos que possuem em conta. Tendo em vista esse cenário, foi feita uma normalização dos dados para equilibrar melhor a escala. A normalização foi feita usando o np.log10 da biblioteca numpy. A normalização de dados usando a função np.log10 transforma os valores de uma variável original em uma escala logarítmica de base 10. Esse tipo de normalização é útil para reduzir a variabilidade e atenuar a influência de valores muito altos, facilitando a análise e visualização dos dados. Após a normalização o boxplot dos dados ficou da seguinte maneira:



Uma base de dados muito mais equilibrada para se extrair correlações válidas.

Após a normalização dos dados, foram removidos usuários com menos de 2 jogos jogados, pois entende-se que esses usuário não contribuem significativamente para a construção de uma correlação entre os jogos e podem influenciar negativamente nos resultados.

Em seguida os dados foras separados em treino e teste na proporção 80/20. É importante que aqui os dados sejam separados de forma que em ambas as bases, tenham os mesmos usuários, para que possamos futuramente comparar os resultados para os usuários e avaliar quantas indicações feitas pelo modelo foram corretas.

Com a base de dados devidamente preparada, utilizamos o conjunto de treinamento para gerar a matriz de correlação entre os jogos, empregando a função corr. Nesse estágio, torna-se evidente o desafio de escalabilidade desse modelo, uma vez que a matriz de correlação gerada possui aproximadamente 10.000 linhas por 10.000 colunas. Esse tamanho impõe um custo computacional significativo, que se torna ainda mais crítico em empresas de grande porte, onde a base de dados é ainda maior, exigindo uma elevada capacidade de processamento.

O método de fatorização de matrizes mencionado na fundamentação teórica mitiga este problema. Essa técnica divide a matriz original em matrizes menores e correspondentes, o que reduz o custo computacional e torna o processamento mais viável para grandes volumes de dados.

Com a matriz de correlação, podemos então buscar um usuário e recomendar seus top 10 jogos recomendados com base nas correlações mais fortes de jogos que ele já possui e mais joga. Com isso geramos as 10 recomendações para o usuário com user_id = "--000--"

	item_name	value	user_id
0	PAYDAY 2	0.230286	--000--
1	Rust	0.211344	--000--
0	Unturned	0.267594	--000--
1	Brawlhalla	0.245879	--000--
0	Robocraft	0.281113	--000--
1	Trove	0.267594	--000--
0	Killing Floor 2	0.326063	--000--
1	Killing Floor Mod: Defence Alliance 2	0.325921	--000--
0	Saints Row IV	0.339459	--000--
1	Saints Row 2	0.263086	--000--

Aqui, optou-se por selecionar os 5 jogos mais jogados de um usuário e, para cada um dos jogos, selecionar as 2 maiores correlações, com isso geramos uma lista de 10 jogos recomendados para o usuário.

A escolha de escolher 5 jogos mais jogados pelo usuário é para que se busque uma variedade de jogos. Se um jogador jogou muito o jogo *Counter-Strike* as correlações mais fortes para esse jogo serão provavelmente outros jogos da franquia *Counter-Strike* e geraria uma lista muito previsível e repetitiva de recomendações. Escolhendo os 5 jogos mais jogados, evita-se esse problema e se constrói um modelo de recomendação mais completo.

4.1 AVALIAÇÃO DO MODELO

Para avaliação dos resultados do modelo, geramos o top 10 de jogos recomendados para cada um dos usuários na base de treinamento, com isso, cruzamos a base gerada com a base de testes para conferir quantas das recomendações geradas pelo modelo configuram jogos que o usuário realmente comprou e jogou. Com isso geramos a razão de jogos indicados/jogos comprados, uma métrica simples e direta de visualização da eficiência do modelo.

Para uma base de dados com aproximadamente 60mil usuários, a acurácia do modelo foi de 10%. Isso quer dizer que, a cada 10 jogos recomendados, 1 jogo foi realmente adquirido. É importante citar também que 57% dos jogares tiveram jogos recomendados corretamente.

5. CONCLUSÃO

A utilização de modelos de recomendação baseados em métricas de correlação demonstrou ser uma abordagem prática e eficiente para cenários que demandam simplicidade e facilidade de implementação. O método mostrou-se capaz de gerar sugestões de itens relacionados, identificando associações relevantes entre jogos com base no comportamento dos usuários.

No entanto, os resultados indicaram limitações significativas em termos de personalização e capacidade de capturar as complexidades das preferências individuais, como evidenciado pela acurácia de 10%. Embora esse desempenho seja adequado para situações específicas, ele ressalta que a abordagem baseada exclusivamente em correlação não é ideal para contextos mais complexos ou para grandes catálogos de produtos, onde é necessária uma compreensão mais profunda do perfil dos usuários.

Apesar disso, o modelo apresenta um ponto de partida sólido, destacando-se pela sua simplicidade e capacidade de gerar recomendações úteis para uma parcela considerável dos usuários. Ele também contribui para o entendimento inicial de associações entre jogos no catálogo da Steam.

Assim, este trabalho reforça a importância de sistemas de recomendação na melhoria da experiência do usuário em plataformas digitais e aponta caminhos para pesquisas futuras que possam tornar essas soluções ainda mais eficazes.

6. TRABALHOS FUTUROS

Embora o modelo desenvolvido tenha demonstrado potencial em recomendar jogos com base em correlações entre itens, os resultados obtidos indicam diversas oportunidades de melhoria que podem ser exploradas em trabalhos futuros.

Uma das principais direções é a análise aprofundada dos hiperparâmetros utilizados no modelo. Decisões como a exclusão de usuários com menos de dois jogos, a quantidade de jogos considerados no top de recomendações, e a seleção dos cinco jogos mais jogados de cada usuário podem ter impacto significativo no desempenho do sistema. Estudos experimentais para ajustar esses parâmetros e identificar suas combinações ideais são necessários para maximizar a acurácia e a relevância das recomendações.

Além disso, a aplicação de modelos mais sofisticados e robustos apresenta uma perspectiva promissora. Abordagens baseadas em filtragem colaborativa, como a fatoração de matrizes (ALS, SVD), ou métodos baseados em aprendizado de máquina, como redes neurais e modelos híbridos, podem capturar padrões mais complexos de comportamento e preferências dos usuários. Esses métodos, embora mais exigentes computacionalmente, têm o potencial de superar as limitações da matriz de correlação, oferecendo maior personalização e escalabilidade em bases de dados maiores e mais diversificadas.

Outro aspecto relevante é o tratamento do problema do "cold start", comum em sistemas de recomendação, que afeta especialmente novos usuários e itens. Métodos que combinem dados explícitos (como avaliações e preferências declaradas) com dados implícitos podem ser explorados para atenuar esse desafio.

Por fim, estudos futuros também podem se concentrar em métricas mais refinadas de avaliação do modelo, buscando capturar não apenas a precisão das recomendações, mas também sua relevância e impacto na experiência do usuário. A integração de

feedback qualitativo dos usuários, por exemplo, pode complementar as análises quantitativas e fornecer uma visão mais abrangente sobre a eficácia do sistema de recomendação.

7. REFERÊNCIAS BIBLIOGRÁFICAS

HU, Yifan; KOREN, Yehuda; VOLINSKY, Chris. **Collaborative Filtering for Implicit Feedback Datasets**. AT&T Labs – Research, Florham Park, NJ, Yahoo! Research, Haifa, Israel.

GERMAIN, Audrey. ***Game Recommendation System***. Repositório GitHub. Disponível em: https://github.com/AudreyGermain/Game-Recommendation-System. Acesso em: 2 out. 2024.

VICTOR. ***ALS Implicit Collaborative Filtering***. Medium, 23 ago. 2017. Disponível em: https://medium.com/radon-dev/als-implicit-collaborative-filtering-5ed653ba39fe. Acesso em: 5 out. 2024.

OMIXDATA. ***Estatística: Análise de correlação usando Python e R. Medium***, 22 abr. 2020. Disponível em: https://medium.com/omixdata/estat%C3%ADstica-an%C3%A1lise-de-correla%C3%A7%C3%A3o-usando-python-e-r-d68611511b5a. Acesso em: 3 nov. 2024.

MCAULEY, Julian. ***Datasets***. University of California, San Diego, 2016. Disponível em: https://cseweb.ucsd.edu/~jmcauley/datasets.html#steam_data. Acesso em: 5 nov. 2024.

ALEXANDER, William***. Steam Game Recommendation System***. Medium, 19 jun. 2021. Disponível em: https://medium.com/@william.alexander23326/steam-game-recommendation-system-875a4912a468. Acesso em: 1 nov. 2024.