Melhorando a Experiência do Usuário: Um Sistema Inteligente de Recomendação de Filmes com Matrix Factorization e KNN

Lara Fernandes Pereira[1], Renan Lopes Silva[1]

[1]Departamento de Informática – Universidade Federal de Viçosa (UFV)

{lara.f.pereira,renan.l.silva}@ufv.br

Resumo. *Projeto final apresentado à disciplina de Inteligência Artificial I, de código INF 420, ministrada pelo professor Julio C. S. Reis, como requisito parcial para aprovação na disciplina.*

1. Introdução

Com a crescente oferta e disponibilidade de conteúdos midiáticos atualmente, encontrar um conteúdo correspondente aos seus interesses pode ser uma tarefa desafiadora para os usuários. Nesse cenário, os sistemas de recomendação desempenham um papel crucial ao filtrar e fornecer sugestões personalizadas aos usuários, ajudando-os a descobrir conteúdos interessantes, considerando seus históricos de avaliação e preferências individuais. Partindo dessa premissa, o presente trabalho explora a construção de um sistema inteligente de recomendação de filmes utilizando os métodos *Matrix Factorization* e KNN.

2. Metodologia

2.1 Matrix Factorization

A Fatoração de Matriz (*Matrix Factorization*) é um método utilizado em sistemas de recomendação que representa itens e usuários na forma de vetores de fatores inferidos a partir dos padrões de classificação dos dados. As recomendações são frutos de altas correspondências entre os fatores de um usuário e um item. (Koren, et al. 2009). Existem diferentes algoritmos de factorização de matriz utilizados em sistemas de recomendação, como SVD (*Singular Value Decomposition*), SVD++ e NMF (*Non-Negative Matrix Factorization*).

O SVD é um método utilizado em sistemas de recomendação capaz de oferecer recomendações mais eficientes ao reduzir a carga computacional. Esse método fornece predições por meio da análise de diversos dados, transformando o usuário e os itens selecionados em um espaço de fatores latentes semelhantes. (Ko, et al. 2022). O SVD++ é uma extensão do SVD que leva em consideração feedback

implícito. Tipos de feedback implícito incluem histórico de compras, histórico de navegação, padrões de pesquisa ou até movimentos do mouse. Por exemplo, um usuário que comprou muitos livros pelo mesmo autor provavelmente gosta desse autor (Koren, Y. 2008).

O NMF também é um tipo de algoritmo de fatoração de matrizes, que fatoriza uma matriz não negativa no produto de dois não negativos. Como uma das principais vantagens é possível citar a rápida convergência.

2.2 K-Nearest Neighbor

O *K-Nearest Neighbor* (KNN) é um algoritmo utilizado para classificar conjuntos de dados com base nos K vizinhos mais próximos. O KNN classifica os dados com base na similaridade entre os itens de dados por meio de medidas de distância, como a distância euclidiana, similaridade de cosseno, correlação de Pearson, entre outros. O algoritmo K-Nearest Neighbor (KNN) é comumente usado em sistemas de recomendação para classificar o padrão de busca do usuário e prever suas preferências futuras através da análise dos registros de atividade do usuário, recomendando itens relevantes. Essa abordagem permite uma personalização mais precisa das recomendações, levando em consideração o comportamento específico de cada usuário. (Ko, et al. 2022).

Foram utilizadas diferentes versões do KNN na elaboração do trabalho além do KNN básico. O *KNN with Means*, que utiliza similaridade entre itens e médias de avaliações. O *KNN with Z-Score*, que utiliza uma medida estatística, chamada de z-score, que descreve a posição relativa de um dado em relação à média e ao desvio padrão de um conjunto de dados. E o KNN Baseline, que incorpora uma linha de base, geralmente representada pela classificação média de todos os itens ou a classificação média de um usuário específico, para melhorar a precisão das previsões em sistemas de recomendação.

2.3 Dataset

O MovieLens é um dos principais conjuntos de dados com avaliações de filmes, utilizado em pesquisas de sistemas de recomendação. O dataset foi construído pela GroupLens Research, da Universidade de Minnesota, e contém avaliações coletadas de usuários reais. Nesse estudo foi utilizado como dataset o “MovieLens 1M”, que possui 1 milhão de avaliações de filmes, com 6000 usuários e 4000 filmes.

2.4 Métrica de Análise

O RMSE *(Root Mean Squared Error*) é uma alteração do cálculo de MSE (*Mean Squared Error*). O MSE calcula a média de diferença entre o valor predito com o real, elevando a diferença ao quadrado a fim de penalizar valores que sejam muito altos. Contudo, esse cálculo altera a unidade de medida. Supondo que o problema em questão utilizasse metros como unidade de medida, o resultado do MSE seria em m^2. (Júnior. 2021)

$$MSE\,(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Figura 1. Equação do erro quadrático médio

O RMSE faz o mesmo cálculo do MSE, entretanto, com o intuito de manter a unidade de medida utilizada no dado original, é calculada a raiz quadrada do resultado, fazendo com que um problema que utiliza metros como unidade de medida possua o resultado dessa métrica também em metros.

$$RMSE\,(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Figura 2. Equação da raiz do erro quadrático médio

O RMSE leva em consideração tanto a magnitude quanto a direção dos erros, dando mais peso a erros maiores, sendo sensível a valores discrepantes. Quanto maior o resultado do RMSE, menos preciso é o modelo.

2.5 Outras ferramentas

Durante o desenvolvimento foi utilizada a biblioteca *Surprise,* que implementa algumas funções do SciKit. A *Surprised* é uma biblioteca de filtragem colaborativa, projetada para facilitar o desenvolvimento e a implementação de algoritmos de recomendação. O objetivo principal da *Surprise* é fornecer uma estrutura simples para realizar avaliações ou recomendações de itens com base nas preferências dos usuários. A biblioteca permite a implementação dos algoritmos citados acima.

3. Discussão e Resultados

O dataset do MovieLens foi desenvolvido para ser utilizado em pesquisas de sistemas de recomendação, portanto não foi necessário realizar nenhum processamento adicional além da importação e agregação das bases de filmes e usuários. O dataset resultante possui colunas referentes ao id do usuários, id do filme e avaliação do usuário. Assim como o MovieLens, o Surprise também foi estruturado para facilitar o processo de recomendação. Devido a isso foi possível aplicar algoritmos de filtragem colaborativa e obter recomendações personalizadas de forma simples.

3.1 Análise do dataset

O gráfico a seguir mostra quantos filmes estão categorizados em cada gênero. Um filme pode se categorizar em mais de um gênero.

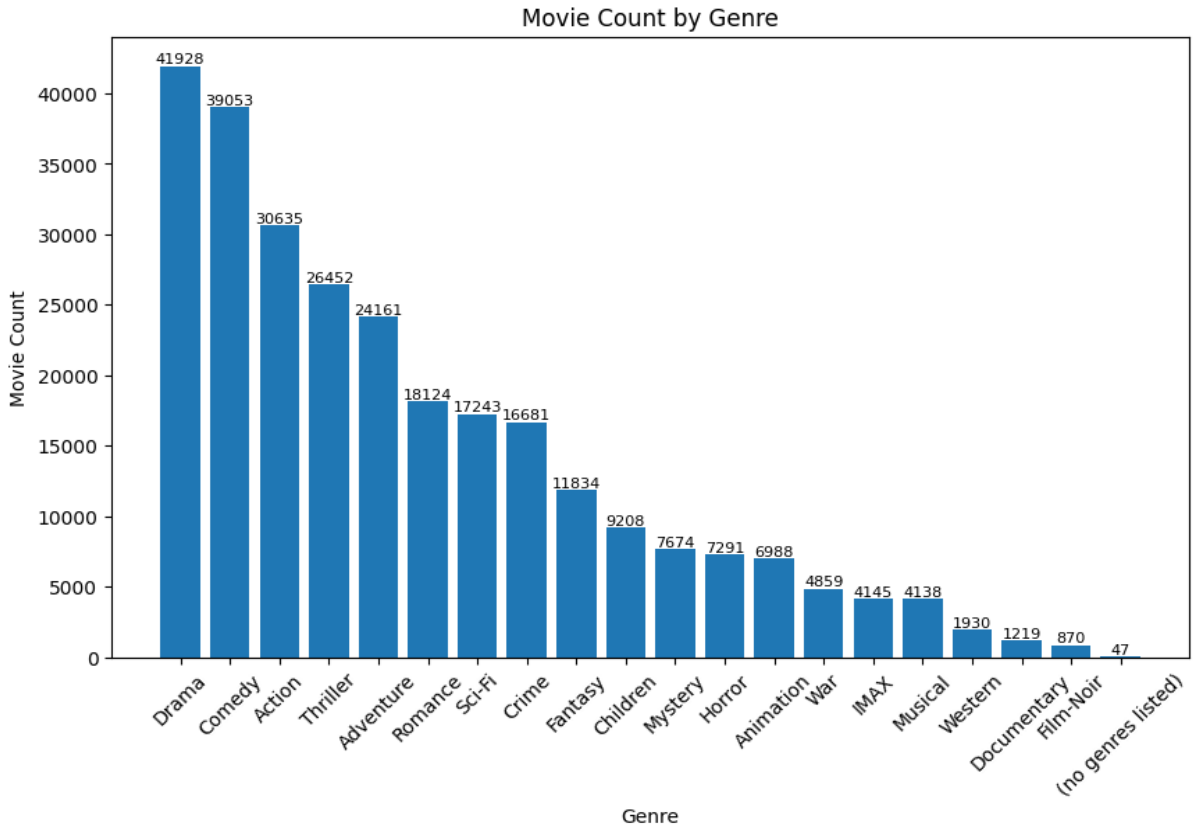

Figura 3. Contagem de filmes por gênero

O gráfico abaixo demonstra como as avaliações dos filmes estão distribuídas no dataset. As notas variam de 0.5 a 5.0, em uma escala de 0,5.

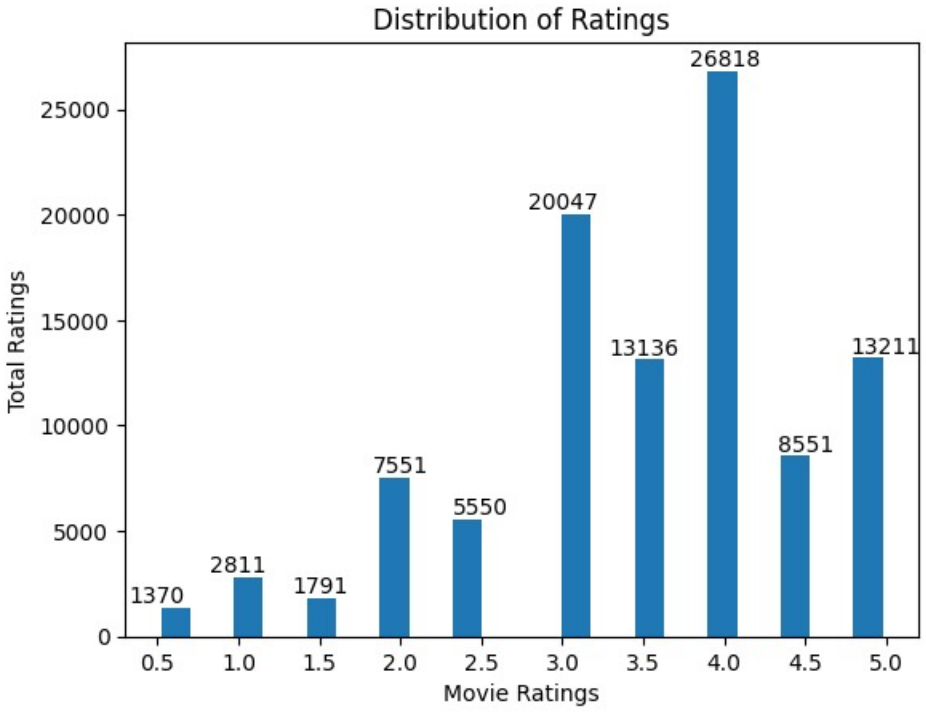

Figura 4. Distribuição das avaliações

A realização dos testes foi feita utilizando como parâmetro base os dados do usuário anonimo que será referido com “usuário X” do conjunto de dados do MovieLens. A tabela a seguir demonstra os 15 primeiros resultados ao ordenar as avaliações desse usuário por nota de forma decrescente.

Tabela 1. Avaliações do usuário 1

Título	Gêneros	Avaliação
M*A*S*H (a.k.a. MASH) (1970)	[Comedy, Drama, War]	5.0
Excalibur (1981)	[Adventure, Fantasy]	5.0
Indiana Jones and the Last Crusade (1989)	[Action, Adventure]	5.0
Pink Floyd: The Wall (1982)	[Drama, Musical]	5.0
From Russia with Love (1963)	[Action, Adventure, Thriller]	5.0
Goldfinger (1964)	[Action, Adventure, Thriller]	5.0
Dirty Dozen, The (1967)	[Action, Drama, War]	5.0
Gulliver's Travels (1939)	[Adventure, Animation, Children]	5.0
American Beauty (1999)	[Drama, Romance]	5.0
South Park: Bigger, Longer and Uncut (1999)	[Animation, Comedy, Musical]	5.0
Austin Powers: International Man of Mystery (1997)	[Action, Adventure, Comedy]	5.0
Face/Off (1997)	[Action, Crime, Drama, Thriller]	5.0
Conan the Barbarian (1982)	[Action, Adventure, Fantasy]	5.0
L.A. Confidential (1997)	[Crime, Film-Noir, Mystery, Thriller]	5.0
Iron Giant, The (1999)	[Adventure, Animation, Children, Drama, Sci-Fi]	5.0

3.2 Resultados Matrix Factorization

O próximo gráfico demonstra o RMSE dos métodos de Fatoração de Matriz utilizados. Como pode ser observado, o NMF obteve maior RMSE e o SVD++ obteve o menor valor. Isso indica que no contexto em que os métodos foram aplicados, o SVD++ obteve melhores resultados, enquanto o NMF obteve resultados mais distantes do que os esperados.

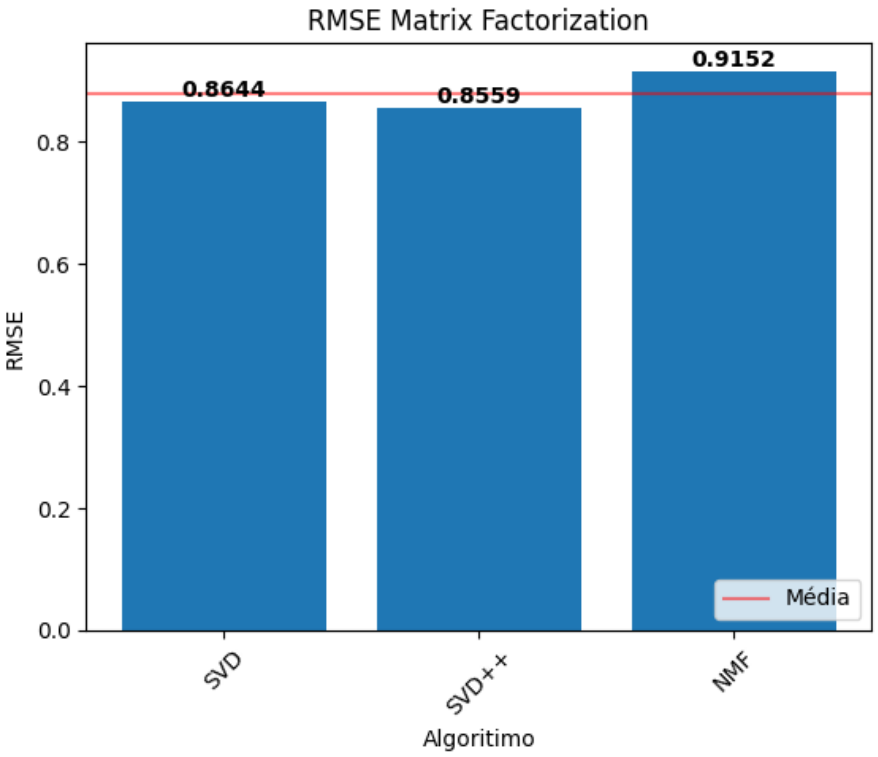

Figura 5. RSME dos métodos de Fatoração de Matriz

As recomendações exibidas por cada método foram:

Tabela 2. Recomendações dos métodos Matrix Factorization

SVD	SVD++	NMF
Goodfellas (1990)	Schindlers List (1993)	Fargo (1996)
Apocalypse Now (1979)	Lock, Stock & Two Smoking Barrels (1998)	Alien (1979)
Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)	The Shining (1980)	Seven (a.k.a. Se7en) (1995)
Monty Python and the Holy Grail (1975)	Goodfellas (1990)	Goldfinger (1964)
Schindlers List (1993)	Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)	Apocalypse Now (1979)
Fargo (1996)	Heat (1995)	Dogma (1999)
The Silence of the Lambs (1991)	The Wizard of Oz (1939)	Schindlers List (1993)
The Shining (1980)	Monty Python and the Holy Grail (1975)	Young Frankenstein (1974)
Office Space (1999)	Easy Rider (1969)	The Silence of the Lambs (1991)
L.A. Confidential (1997)	Toy Story (1995)	Lock, Stock & Two Smoking Barrels (1998)

3.3 Resultados KNN

O próximo gráfico demonstra o RMSE dos diferentes métodos KNN utilizados. Houve maior semelhança entre os resultados dos KNN do que entre os métodos demonstrados acima. O método KNN Baseline foi o que obteve o melhor resultado, enquanto o KNN padrão forneceu resultados mais distantes do esperado do que os outros métodos.

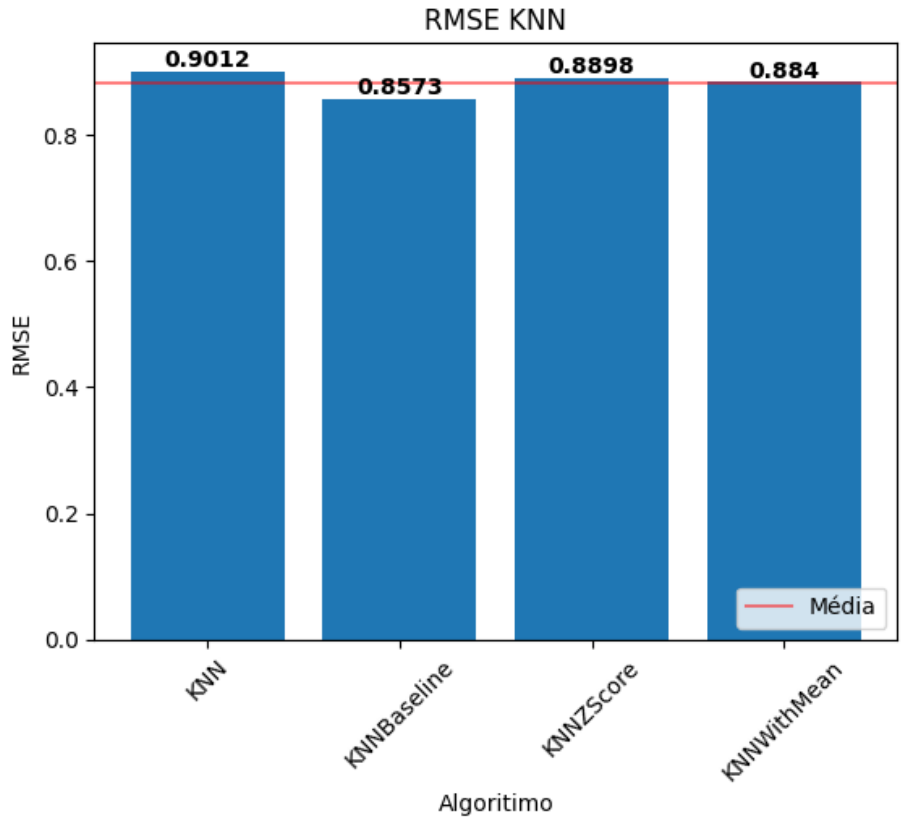

Figura 6. RSME dos métodos KNN

As recomendações exibidas por cada método foram:

Tabela 3. Recomendações dos métodos KNN

KNN Basic	KNN Baseline	KNN With Means	KNN With Zscore
Rushmore (1998)	Princess Bride (1987)	The Princess Bride (1987)	Pulp Fiction (1994)
Forrest Gump (1994)	Forrest Gump (1994)	All Quiet on the Western Front (1930)	The Princess Bride (1987)
Braveheart (1995)	Monty Python and the Holy Grail (1975)	Forrest Gump (1994)	Forrest Gump (1994)
The Princess Bride (1987)	Pulp Fiction (1994)	Raiders of the Lost Ark (1981)	Monty Python and the Holy Grail (1975)

Duck Soup (1933)	Raiders of the Lost Ark (1981)	Pulp Fiction (1994)	Big Lebowski (1998)
Big (1988)	Saving Private Ryan (1998)	Monty Python and the Holy Grail (1975)	Tommy Boy (1995)
Platoon (1986)	Braveheart (1995)	Saving Private Ryan (1998)	Raiders of the Lost Ark (1981)
Back to the Future (1985)	Back to the Future (1985)	Platoon (1986)	Braveheart (1995)
Monty Python and the Holy Grail (1975)	A Clockwork Orange (1971)	Braveheart (1995)	Duck Soup (1933)
Charlottes Web (1973)	Platoon (1986)	Back to the Future (1985)	A Clockwork Orange (1971)

3.4 Resultados gerais

O RMSE dos algoritmos utilizados demonstra que, entre todos os modelos aplicados, o NMF e o KNN obtiveram as menores precisões, enquanto o KNN Baseline e o SVD++ foram os mais precisos, tendo em vista que a precisão de um modelo é inversamente proporcional ao valor do seu RMSE.

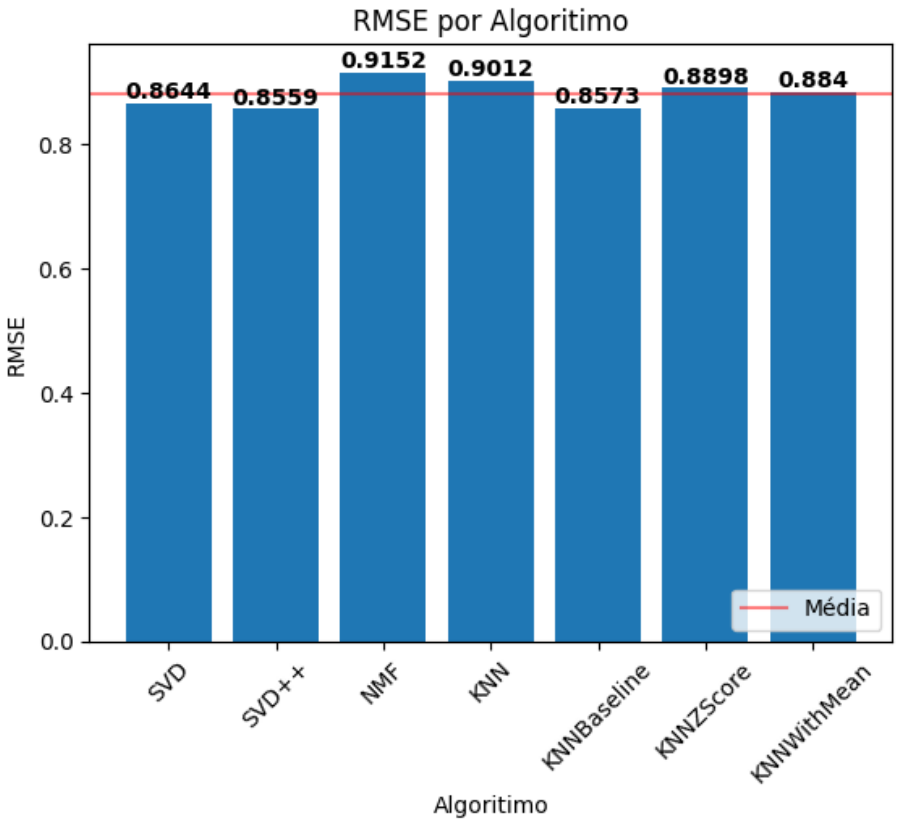

Figura 7. RSME dos métodos Matrix Factorization e KNN

O método do SVD++ obteve menor RMSE. Os a seguir demonstram, respectivamente, os gêneros dos 30 filmes mais bem avaliados pelo Usuário X e as 30 recomendações de filmes mais próximas realizadas pelo SVD++.

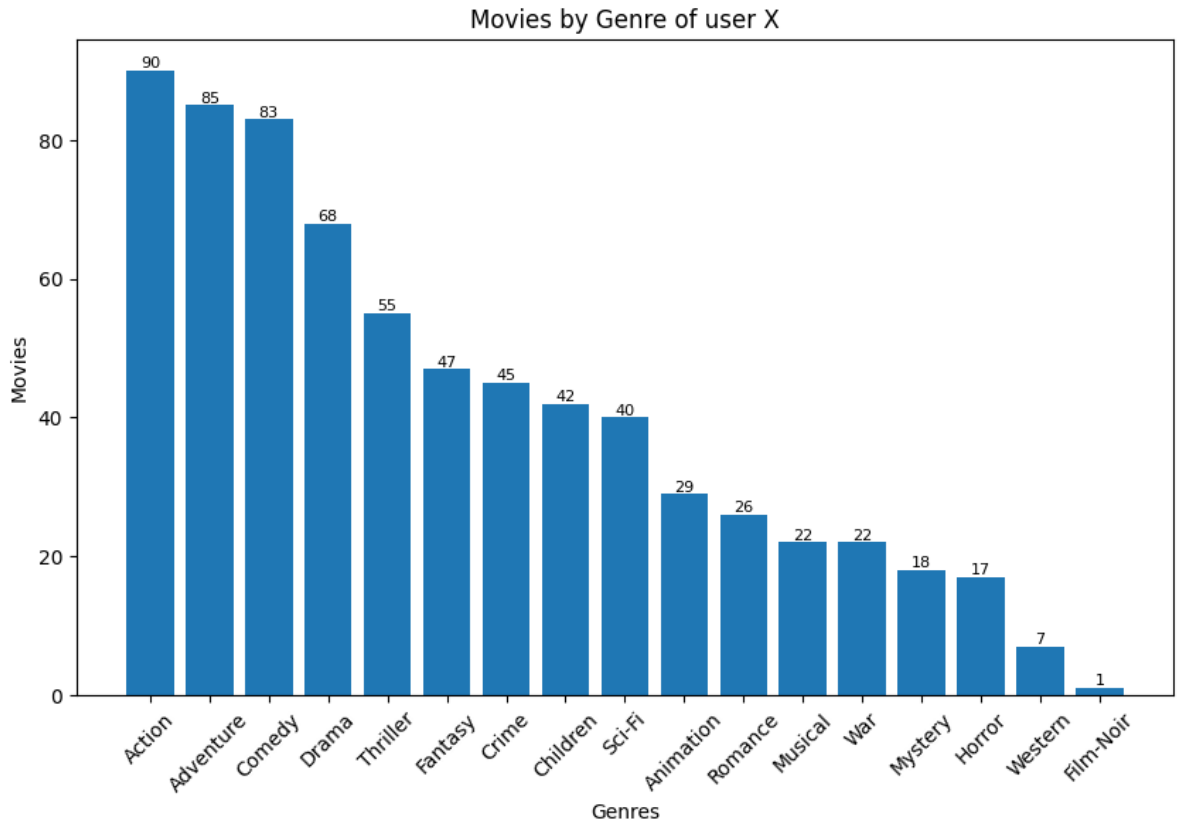

Figura 8. Distribuição dos gêneros dos filme assistidos pelo usuário X

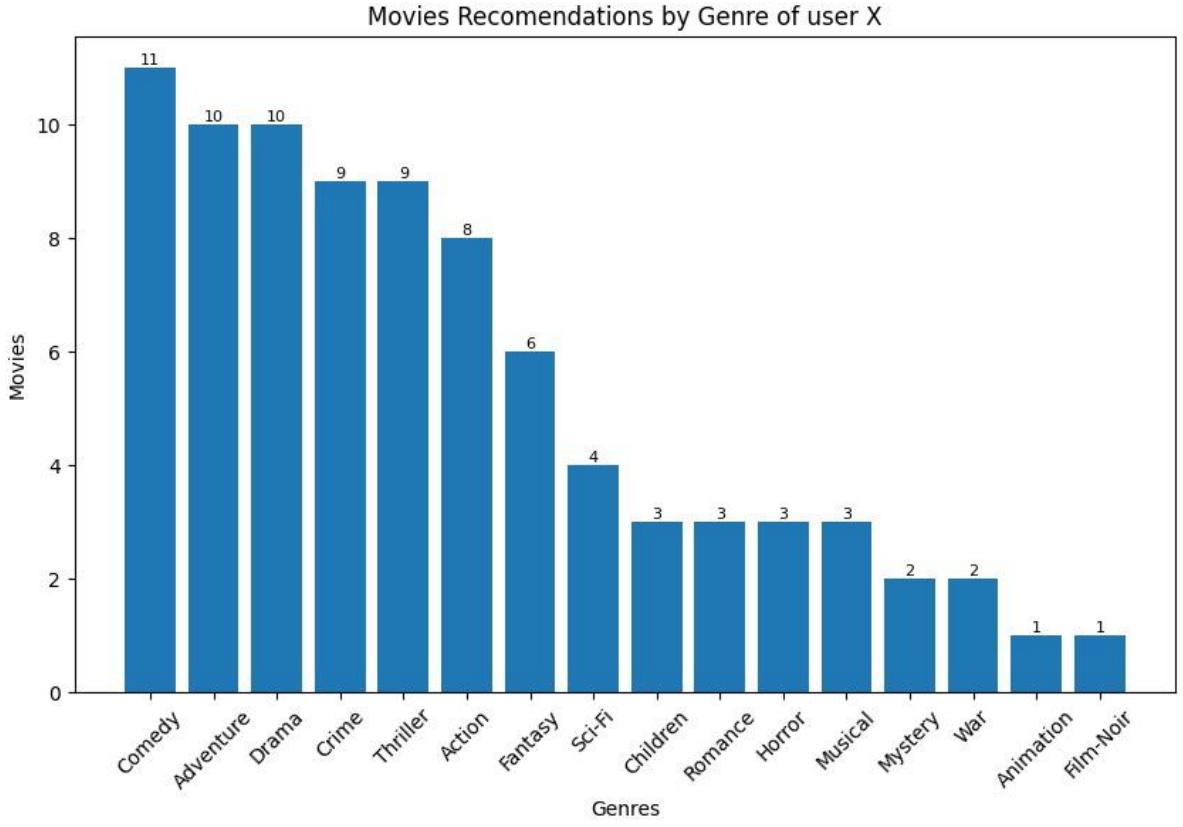

Figura 9. Distribuição dos gêneros dos filmes recomendados pelo SVD++

4. Código

A implementação do presente trabalho está disponível no repositório https://github.com/laracolorida/movie-match .

5. Conclusão

Em conclusão, esse trabalho investigou a eficácia de diferentes algoritmos de recomendação, incluindo SVD, SVD++, NMF, KNN, KNN Baseline, KNN With Mean e KNN Z-Score. A análise dos resultados revelou que tanto o SVD++ quanto o KNN Baseline apresentaram desempenhos semelhantes em termos de precisão na geração de recomendações ao se analisar os resultados do RMSE referente a cada um deles. Com base nos resultados obtidos, podemos concluir que ambos são algoritmos de recomendação eficazes, capazes de fornecer recomendações personalizadas aos usuários.

É importante ressaltar que, apesar dos resultados promissores, este estudo possui algumas limitações. Dentre elas, destacam-se as restrições de hardware, o que acarretou na ausência de de um dataset atual, pois devido a isso os dataset mais recentes do MovieLens não puderam ser utilizados, tendo em vista que as versões mais recentes possuem mais dados e por consequência exigindo maior poder computacional, resultando na utilização do MovieLens 1M que possui filme até o ano de 2003.

Encorajamos que pesquisas futuras explorem outras abordagens e considerem a combinação de múltiplos algoritmos, além de uma abordagem que utilize deep learning para melhorar ainda mais a qualidade das recomendações em sistemas de recomendação.

Referências

F. Maxwell Harper, Joseph A. Konstan. (2015) “The MovieLens Datasets: History and Context”. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4: 19:1–19:19.

JÚNIOR, Clébio de Oliveira. (2021) “Métricas para Regressão: Entendendo as métricas R², MAE, MAPE, MSE e RMSE”. Data Hackers.

Koren, Y. (2008) “Factorization meets the neighborhood “.Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08.

Ko, Hyeyoung, et al. (2022) “A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields”. Electronics, vol. 11, no 1, janeiro de 2022, p. 141.

Koren, Yehuda, et al. (2009) “Matrix Factorization Techniques for Recommender Systems”. Computer, vol. 42, no 8, agosto de 2009, p. 30–37.

Koren, Y., Bell, R., & Volinsky, C. (2009) “Matrix Factorization Techniques for Recommender Systems”. Computer, 42(8), 30–37.

Li, T., Gao, C., & Du, J. (2009) “A NMF-Based Privacy-Preserving Recommendation Algorithm”. 2009 First International Conference on Information Science and Engineering.

Rapaport, Elad. (2022) “MovieLens-1M Deep Dive — Part I: A hands-on recommendation systems tour using the popular benchmark dataset”. Towards Data Science.