




[Marcar como feito](#) [Descrição](#) [Enviar](#) [Editar](#) [Visualizar envios](#) **Data de entrega:** segunda-feira, 30 jun. 2025, 23:55 **Arquivos requeridos:** Notebook.ipynb ( [Baixar](#)) **Número máximo de arquivos:** 2**Tipo de trabalho:**  Trabalho individual

Quarto trabalho - Cluster

O conjunto de dados MovieLens 1M é um conjunto popular para análise e pesquisa em sistemas de recomendação de filmes. Ele foi criado pela GroupLens Research, um grupo de pesquisa da Universidade de Minnesota. O conjunto contém 1000.000 avaliações de filmes feitas por cerca de 6000 usuários sobre +/- 3000 filmes. As avaliações estão em uma escala de 1 a 5, e cada usuário avaliou pelo menos 20 filmes.

Os dados estão divididos em cinco arquivos de texto principais:

1. u.data: contém as avaliações dos usuários, cada linha representa uma avaliação no formato `user_id | movie_id | rating | timestamp`.
2. u.item: contém informações dos filmes, incluindo o título, data de lançamento, gênero e outros detalhes.
3. u.user: contém informações demográficas sobre os usuários, como idade, sexo, ocupação e CEP.
4. u.genre: contém uma lista de gêneros de filmes.
5. u.occupation: contém uma lista de ocupações dos usuários.



Não iremos utilizar os arquivos 3, 4 e 5. Você pode obter o conjunto de dados MovieLens 1M no seguinte link:

- <https://files.grouplens.org/datasets/movielens/>

Baixe o **README** e o arquivo **ml-1m.zip**

Inclua no seu trabalho o comando para baixar esses dois conjuntos de dados e guardá-los em memória. Você pode salvar uma cópia local, mas deve incluir o comando para baixá-los da internet (e comentar esse comando).

O trabalho deve ser entregue como um notebook Jupyter, em Python.

Sigam as instruções abaixo:

1. Importe as bibliotecas necessárias:
2. Crie seções no notebook para organizar cada etapa do exercício:
 - Análise exploratória dos dados:
 - Verifique as informações básicas, como quantidade de filmes, usuários e avaliações.
 - Utilize gráficos para visualizar a distribuição das avaliações, gêneros e quantidade de filmes por ano.
 - Pré-processamento dos dados:
 - Verifique se há algum valor duplicado
 - Normalize os dados, se necessário.
 - Cálculo das matrizes de distâncias
 - Aplicação dos algoritmos de clusterização
 - K-means
 - DBSCAN
 - Clustering hierárquico (*Agglomerative* ou *Divisive*)
 - Análise da Variação do número de clusters
 - Remoção de Outliers (defina critérios)
 - Redução de dimensionalidade

?

- Interpretação dos clusters
- Comparação dos resultados
- Conclusão

3. Use células de texto (Markdown) para descrever cada etapa e, em seguida, células de código para implementar cada etapa.

4. Ao criar gráficos, certifique-se de incluir rótulos e legendas adequadas para facilitar a compreensão.

5. Ao analisar os resultados, use células de texto para discutir suas descobertas e conclusões.

6. Ao final do notebook, escreva uma conclusão geral, destacando as principais descobertas, desafios e possíveis melhorias.

7. Antes de entregar, verifique se todas as células foram executadas na ordem correta e que o notebook está bem organizado.

8. Realize a análise com e sem redução da dimensionalidade:

- Sem redução de dimensionalidade: Utilize a matriz de distâncias pré-calculada.
- Com redução de dimensionalidade: Aplique PCA e SVD nos dados e analise os resultados.
- Plote os valores da matriz diagonal da decomposição SVD.

Interprete os resultados obtidos em cada clusterização e identifique alguns clusters com base nos filmes, como filmes de ficção científica ou de terror.

[VPL](#)

