Projeto Aplicado II

Titulo:

COLLAM FILMS

Sistema de Recomendação de Filmes

AGENDA

01 Informações

02 Apresentação da Empresa

03 Objetivo

04 Dataset

05 Análise Exploratória

06 Limpeza e preparação dos Dados

07 Aprendizado de Máquina

08 Sistema de Recomendação

COLLAM FILMS

Informações

Curso: Tecnologia em Ciências de Dados
Semestre: 3º
Componente curricular: Projeto Aplicado II
Professor: Anderson Adaime de Borba
Integrantes e TIA:

- **Adrieli Machado Zaluski - 22503668**
- **Caroline Ribeiro Ferreira - 22514635**
- **Lais César Fonseca - 22500790**
- **Liliane Gonçalves de Brito Ferraz - 22501142**
- **Múcio Emanuel Feitosa Ferraz Filho - 22515925**
- **Otavio Bernardo Scandiuzzi - 22511921**

COLLAM FILMS

APRESENTAÇÃO DA EMPRESA

A origem do nome da empresa **“COLLAM FILMS”,** nasceu da paixão por filmes e da necessidade de **tornar a experiência** de assistir **filmes** ainda mais **cativantes**. Seu nome é uma fusão das iniciais dos nomes dos integrantes do grupo que deram vida a essa iniciativa, representando nossa colaboração e dedicação.

O nome “Collam” é uma celebração da união e a diversidade de habilidades que empregamos nesse projeto.

COLLAM FILMS

Objetivos do Projeto

1

Melhorar a experiência do usuário ao fornecer recomendações personalizadas com base em seu histórico de filmes assistidos.

2

Criar um modelo eficaz de recomendação de filmes utilizando técnicas de aprendizado de máquina e análise estatística preditiva

3

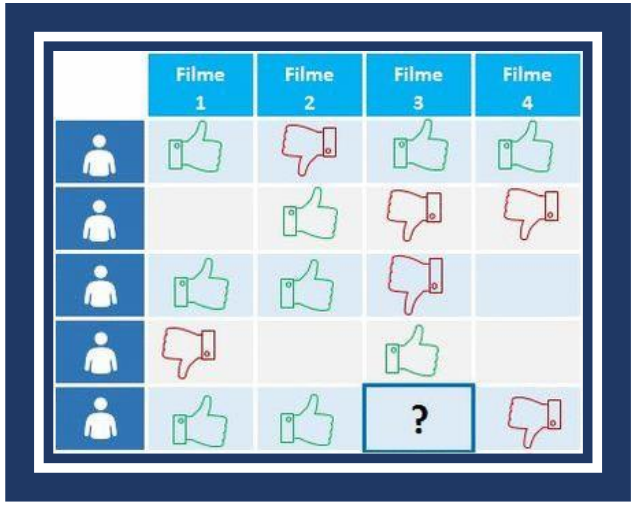
Reduzir o tempo de escolha do próximo filme a assistir, com as recomendações geradas pelo sistema, com base no filme que o usuário apresentar como favorito(s).

COLLAM FILMS

Dataset

Como base de dados para desenvolvimento do projeto e treinamento dados públicos, da plataforma **Kaggle**.

Conjunto de dados **Full MovieLens** - Metadados de Filmes é uma coleção de informações abrangentes sobre filmes lançados até julho de 2017. Ele inclui metadados detalhados sobre aproximadamente **45.000 filmes**, oferecendo uma riqueza de informações relacionadas à indústria cinematográfica.



A base contem também **26 milhões de avaliações de 270 mil usuários** para todos os 45 mil filmes. As classificações estão em uma escala de notas de 1 a 5, obtidas no site oficial do **GroupLens**.

COLLAM FILMS

Análise Exploratória

COLLAM FILMS

ANÁLISE EXPLORATÓRIA

Durante o **processo de exploração** foi possível conhecer melhor os nossos dados através do uso de **métodos estatísticos**, para isso foram seguidos os seguintes passos:

Importação dos pacotes “Pandas”, “Seaborn”, “Numpy” e “Matplotlib”;

Verificações gerais: colunas, tipo de cada coluna, quantidade de dados, valores únicos, valores nulos, ruídos nos dados e descrição geral usando “describe()”;

Separação de variáveis numéricas e categóricas;

Criação de gráficos para uma melhor visualização

COLLAM FILMS

ANÁLISE EXPLORATÓRIA

Análise geral da base de dados, aplicando métodos estatísticos:

	Contagem	Média	Desvio Padrão	Mínimo	25%	50%	75%	Máximo
Receita	45.460	1,12	6,43	0	0	0	0	2,79
Tempo de execução	45.203	9,41	3,84	0	85	95	107	1,26
Votação média	45.460	5,62	1,92	0	5	6	6,8	1,00
Contagem de votos	45.460	1,10	4,91	0	3	10	34	1,41

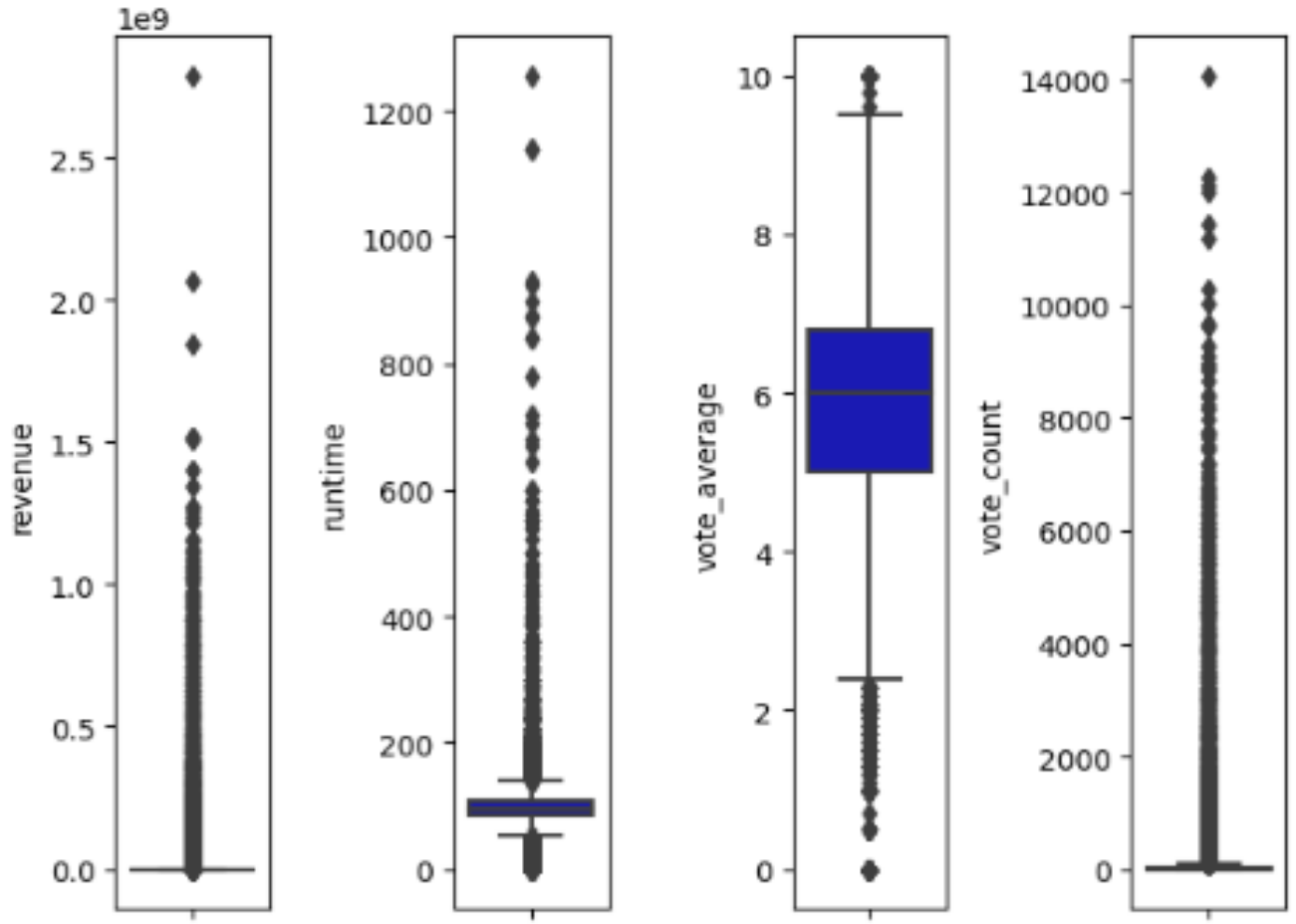
Figura 4: Discribe Dataset movies_metadata.csv

COLLAM FILMS

ANÁLISE EXPLORATÓRIA

Alguns dos gráficos criados para a exploração:

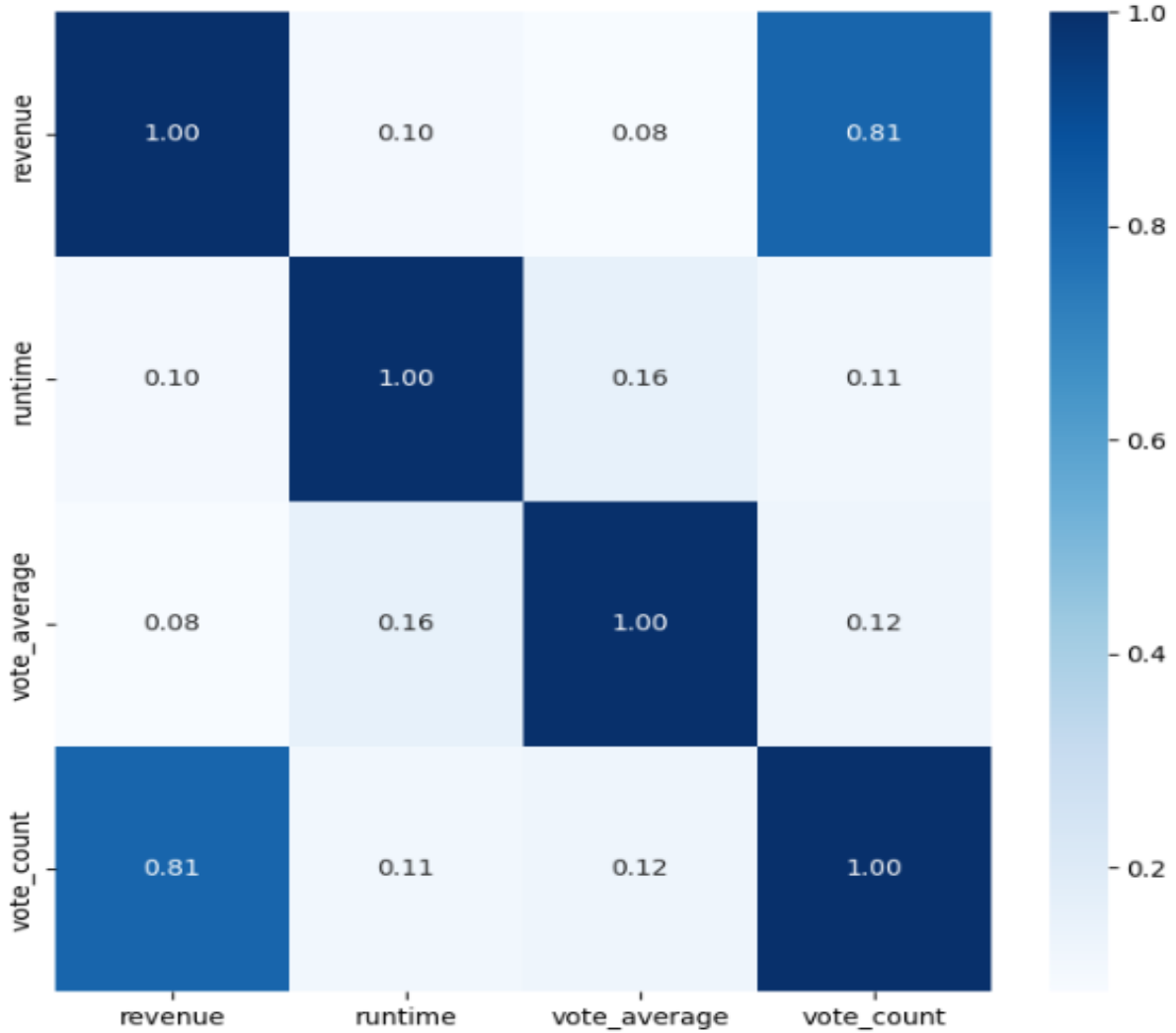
Boxplots das variáveis numéricas:



COLLAM FILMS

ANÁLISE EXPLORATÓRIA

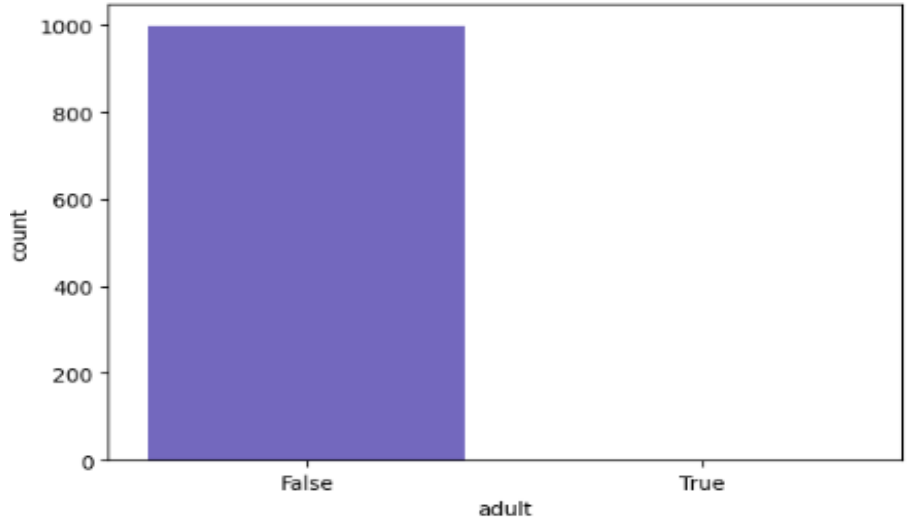
Correlação entre as variáveis numéricas:

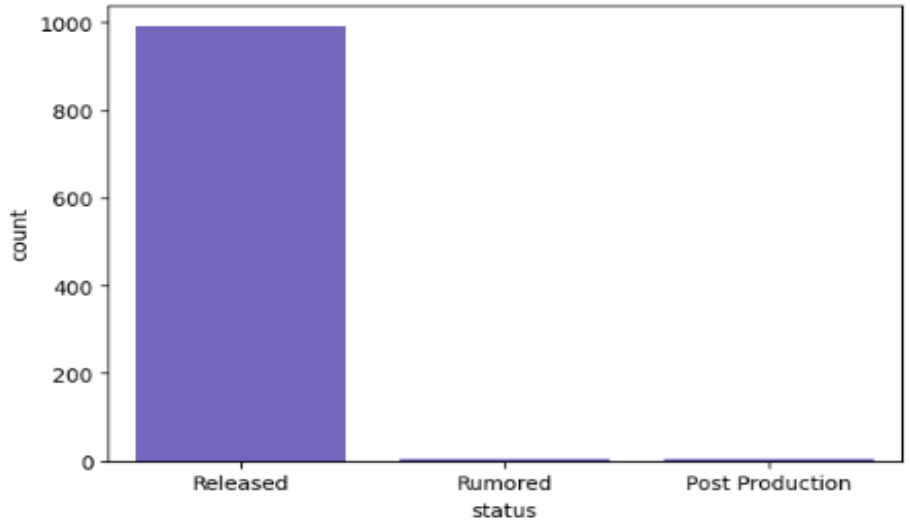


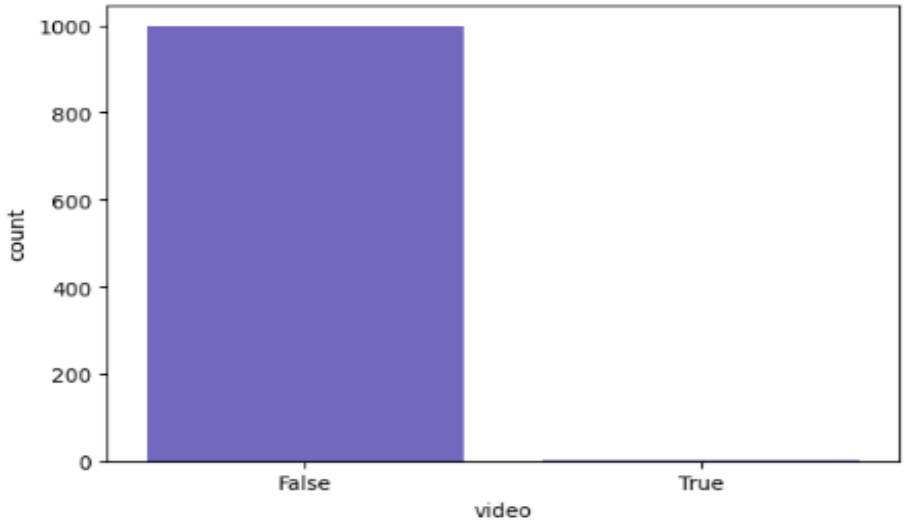
COLLAM FILMS

ANÁLISE EXPLORATÓRIA

Distribuição das variáveis categóricas:







Limpeza e preparação dos Dados

COLLAM FILMS

Limpeza e preparação de dados

Bibliotecas Python
Importar pacotes indicados.

Coletar dados
Importar o dataset a ser analisado

Análise
Entender como os dados estão estruturados

Limpeza
Tratar os dados ausentes e outliers

Estruturar
Definir os dados que serão trabalhados

Sistema de recomendação
Implementar algoritmo

Figura 14: Processo de tratamento de dados

COLLAM FILMS

Limpeza e preparação de dados

1° Filtramos apenas as colunas necessárias de ambos datasets, e as renomeamos.

2° Verificamos os valores nulos, e como foram poucos foram removidos, por não apresentarem impacto nos resultados.

3° Selecionamos apenas usuários que fizeram acima de 400 avaliações.

4° Selecionamos apenas avaliações de usuários que avaliam acima de 400 vezes

5° Selecionamos os filmes que tiveram quantidade de avaliação superior a 30 avaliações.

6° Selecionamos apenas os filmes na linguagem English.

7° Convertemos a variável ID_filme em inteiro.

8° Concatenamos os dataframes.

9° Descartamos os valores duplicados, para não ter o mesmo usuário avaliando o mesmo filme mais de uma vez.

COLLAM FILMS

Sistema de Recomendação

COLLAM FILMS

sistema de Recomendação

Modelo de Sistema de Recomendação

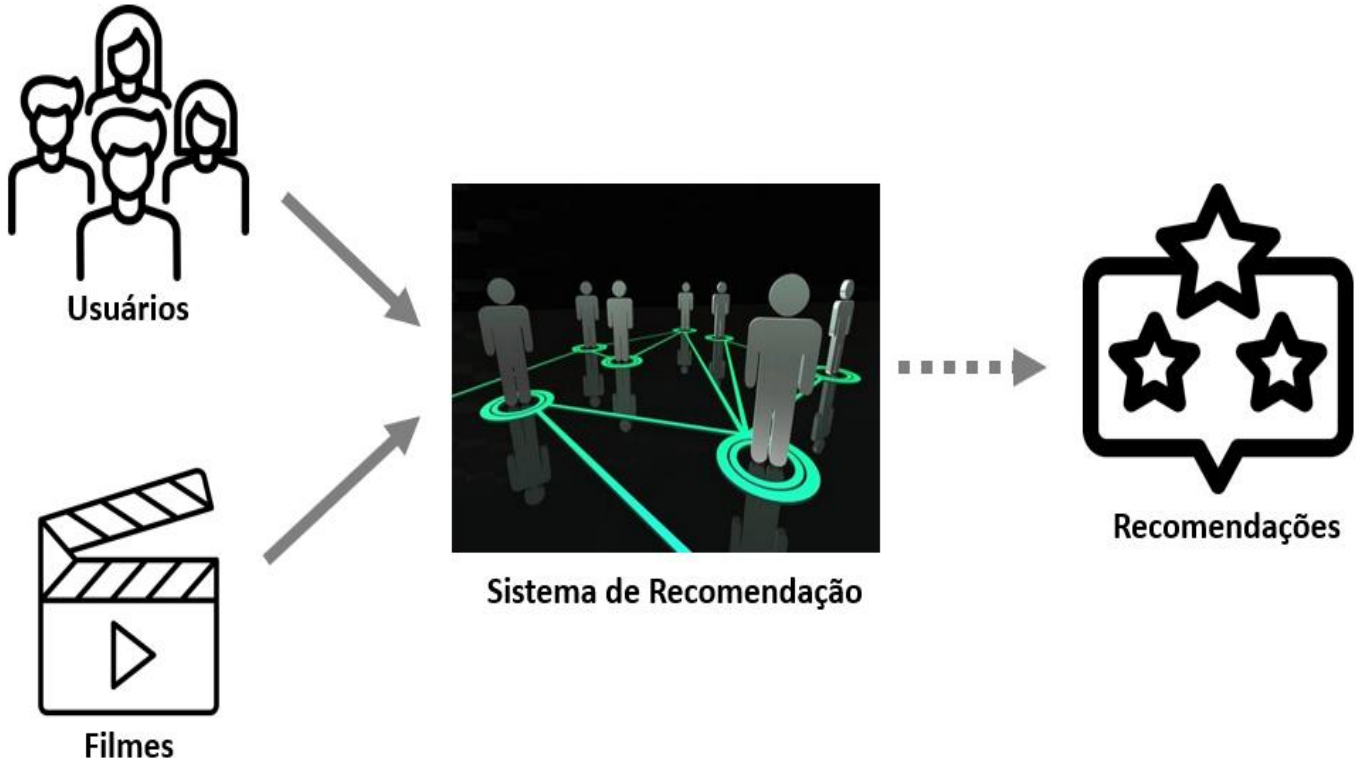


Figura 2: Modelo de Sistema de Recomendação

COLLAM FILMS

Sistema de Recomendação

Utilizamos o sistema de recomendação para alcançar nossos objetivos com este projeto. As principais características de um sistema de recomendação são:

Uso de diferentes Algoritmos	Escalabilidade	Adaptabilidade	Atualização dinâmica	Feedbacks dos usuários
Como a filtragem colaborativa, filtragem baseada em conteúdo e até técnicas de redes neurais	Necessário para fornecer resultados de forma eficiente à medida que o número de componentes aumenta	Pois se adaptam aos interesses de cada usuário.	Importante devido às recorrentes mudanças nos interesses e comportamentos dos usuários	É comum a incorporação de feedbacks implícitos ou explícitos por parte dos usuários

COLLAM FILMS

APRENDIZADO DE MÁQUINA

Filtragem colaborativa X Filtragem colaborativa em conteúdo

Filtragem colaborativa

Utiliza as informações sobre os comportamentos do usuário e verifica se usuários semelhantes gostaram de determinado item para fazer as recomendações.

Filtragem baseada em conteúdo

Recomendações são feitas baseadas em itens semelhantes aos que o usuários gostou previamente. Sendo assim, considera o perfil de determinado usuário e as características dos itens bem avaliados por este

COLLAM FILMS

APRENDIZADO DE MÁQUINA

Escolhemos o método de distancia Euclidiana, para o nosso sistema de recomendação:

Distância Euclidiana

Utilizamos o sistema de recomendação para alcançar nossos objetivos com este projeto. **A distância euclidiana** foi a escolhida para analisar a similaridade entre os filmes e sua aplicação passou pelas seguintes etapas:

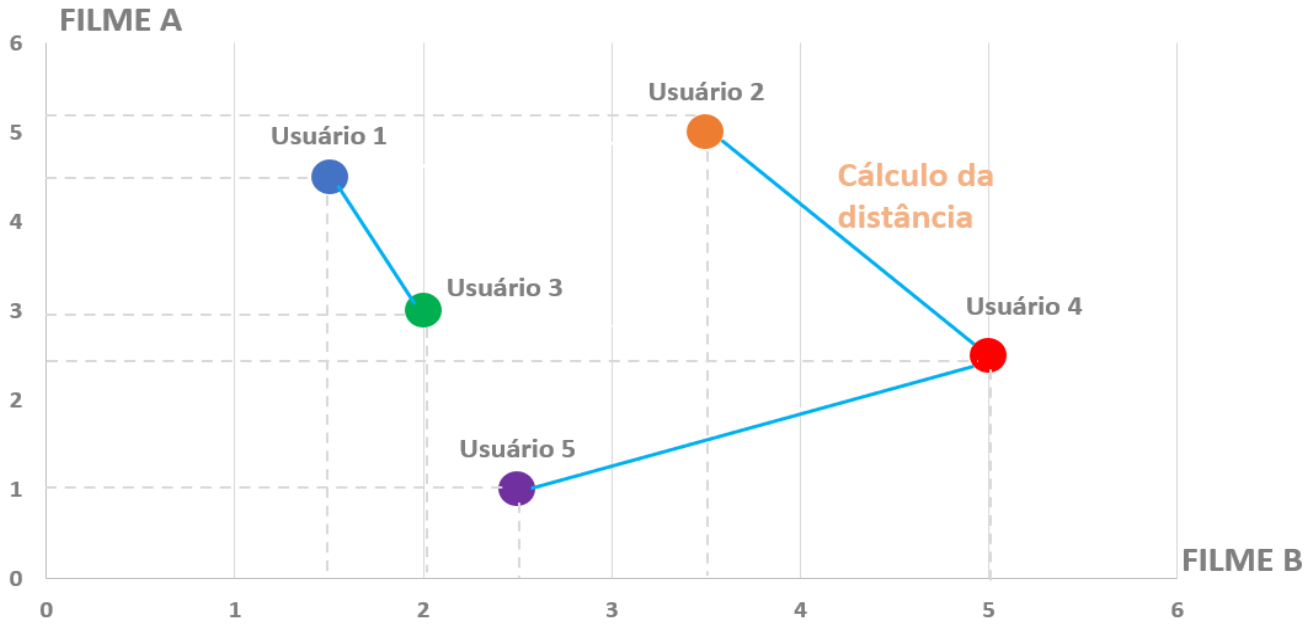


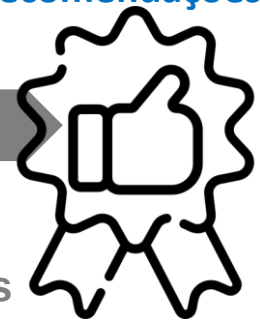
Figura 15: Distância Euclidiana

COLLAM FILMS

Etapas da Recomendação

1. Função para calcular a distância euclidiana entre duas listas
2. Função para calcular a similaridade entre usuários
3. Função para calcular a similaridade entre gêneros
4. Filtrando o dataframe para excluir o filme de referência
5. Calculando a similaridade entre os gêneros dos filmes
6. Combinando as similaridades e Ordenando os filmes por similaridade

Retorna as principais recomendações



Avaliação de Similaridade → **Pontuação de Similaridade**

Combina as similaridades com pesos de 70% para similaridade de usuários e 30% para similaridade de gêneros.

COLLAM FILMS

TESTE E ACURÁCIA

O teste é útil em um aprendizado de máquina para colocar em prático o modelo criado e checar o seu funcionamento. Já a acurácia serve para medir a capacidade deste funcionamento, classificando o seu desempenho.

Uma das medidas de acurácia mais utilizadas em sistemas de recomendações com classificações de usuários é o “Erro Médio Absoluto” (Mean Absolute Error - MAE), Que mede a diferença absoluta entre as classificações previstas e as reais.

COLLAM FILMS

TESTE E ACURÁCIA

As principais características do Mean Absolute Error são:

É calculado pela média das diferenças absolutas entre os valores previstos e o valor real;	É simples e intuitivo para avaliar a precisão de modelos de regressão;	Apresenta baixa sensibilidade a grandes erros individuais quando comparado com outras métricas;	Se apresenta na mesma unidade de medida que os dados de referência

COLLAM FILMS

Obrigado(a)!