TRABALHO PRÁTICO 1 – IA

Grupo: Eduardo Monteiro, Mateus Nunes

1- Definição do problema

Nossa escolha foi de um algoritmo capaz de recomendar filmes através de uma análise do perfil do usuário, ou seja, através de dados fornecidos, descrevendo filmes que o usuário gostou, são geradas recomendações exclusivas que estabelecem uma relação com os já assistidos e logo ,possivelmente, serão de agrado do usuário.

Trata-se de um tema de muito destaque devido a sua ampla aplicação em diversos sistemas, justamente por atrair ainda mais atenção do usuário com uma recomendação embasada e personalizada. Muitos serviços de streaming utilizam de algoritmos com funções semelhantes, como Netflix, Disney+ e diversos outros.

2- Coleta e pré-processamento dos dados

Como nossa base de dados utilizamos o vasto catálogo de filmes presente no site do IMDB, um dos principais portais de avaliação e catalogação de filmes. A base de dados utilizada inicialmente continha mais de 30 mil filmes, além de diversas colunas de características que posteriormente foram retiradas no processo de pré-processamento, como premiações, arrecadação e orçamento.

No Pré-processamento foram adotadas as seguintes mudanças:

- Foram mantidas apenas as colunas de interesse:

```
colunas_interesse = ['title', 'year', 'rating_imdb', 'genre', 'language']
```

- Foram removidos os valores ausentes.

- A coluna ‘genre’ foi transformada em uma lista de gêneros para melhor manipulação dos dados ao longo do algoritmo.

- Foram criados os gráficos a partir dos dados refinados pelo pré-processamento.

3- Escolha da técnica

Foi escolhido o aprendizado não supervisionado, pois não sabemos exatamente qual será a saída do nosso recomendador e da mesma forma não temos rótulos explícitos que indiquem quais filmes o usuário gostou ou não gostou, temos apenas uma entrada de dados com filmes que foram assistidos por esse usuário. Logo temos que definir padrões através dos dados que temos acesso.

Além disso, utilizamos o MaxEclat por ser uma versão modificada do ECLAT com foco em encontrar conjuntos maximais, conjuntos frequentes que nos permitem traçar um perfil para o usuário e através desse perfil cruzar os dados da base de dados completa e gerar recomendações embasadas ao usuário. Outra característica importante é a sua eficiência em bases medias/grandes pois utiliza da interseção de transações ao invés de percorrer toda a base de dados diversas vezes.

4- Desenvolvimento do Modelo

Criação das transações:

- Cada transação corresponde ao Itemset de um filme.

Execução do MaxEclat:

- O algoritmo MaxEclat percorre as transações buscando **conjuntos maximais de itens frequentes**, respeitando um limite mínimo de suporte (por padrão, 0,5%).

Perfil do usuário:

- O usuário fornece um arquivo CSV com seus filmes assistidos.
- O mesmo pré-processamento é aplicado para extrair o Itemset de cada filme assistido.
- O perfil do usuário é construído como a união de todos os seus Itemsets.

Recomendações:

- São selecionados os conjuntos maximais frequentes que mais se sobrepõem com o perfil do usuário.
- Para cada conjunto relevante, são sugeridos filmes que contenham esse conjunto, ainda não assistidos pelo usuário, priorizando maior avaliação (rating_imdb).

5- Análise dos Resultados

A saída do código fornece ao usuário 4 tabelas contendo 5 filmes cada. Cada uma dessas tabelas possui um valor de afinidade próprio, ou seja, quanto maior o valor maior a relação dos filmes recomendados na tabela com os filmes assistidos pelo usuário.

As tabelas apresentam as colunas de características dos filmes recomendados, com nome, ano, nota, diretor, elenco principal e linguagem.

Suponha que o perfil do usuário contenha os seguintes elementos:

['Christopher Nolan', 'Drama', 'Sci-Fi', 'Leonardo DiCaprio', 'Hans Zimmer']

O modelo pode identificar um conjunto frequente maximamente representativo como:

['Drama', 'Sci-Fi', 'Christopher Nolan']

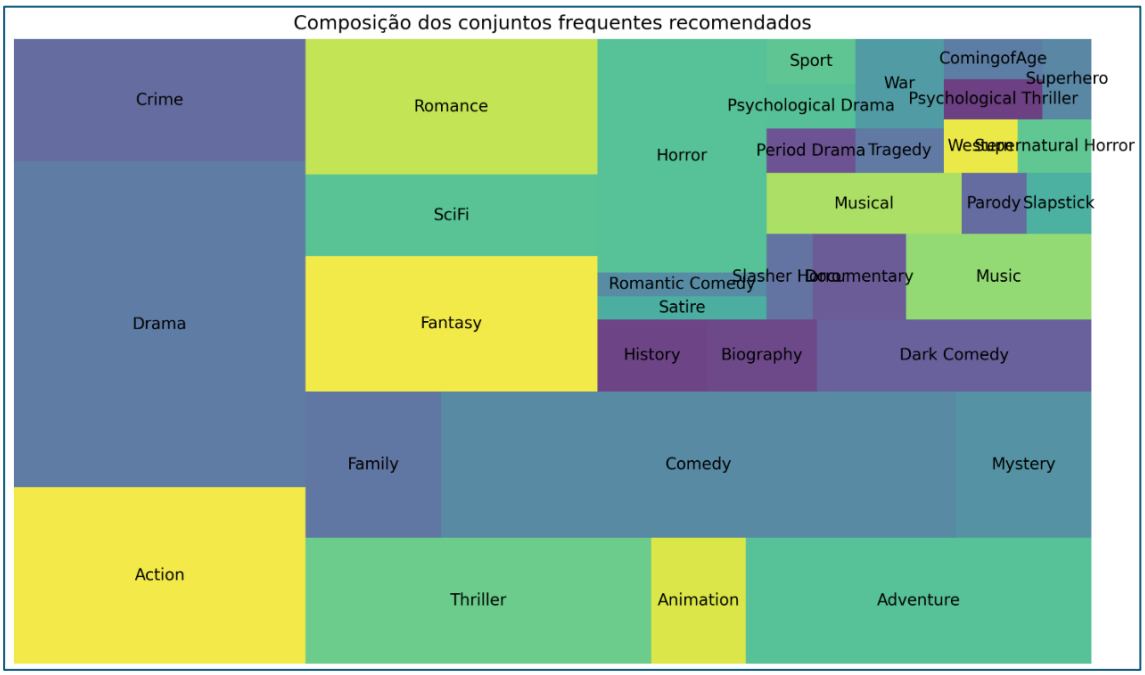
E então recomendar filmes com esse conjunto, como:

Title	**Year**	**IMDb Rating**	**Diretor**
Interstellar	2014	8.6	Christopher Nolan
Inception	2010	8.8	Christopher Nolan
The Prestige	2006	8.5	Christopher Nolan

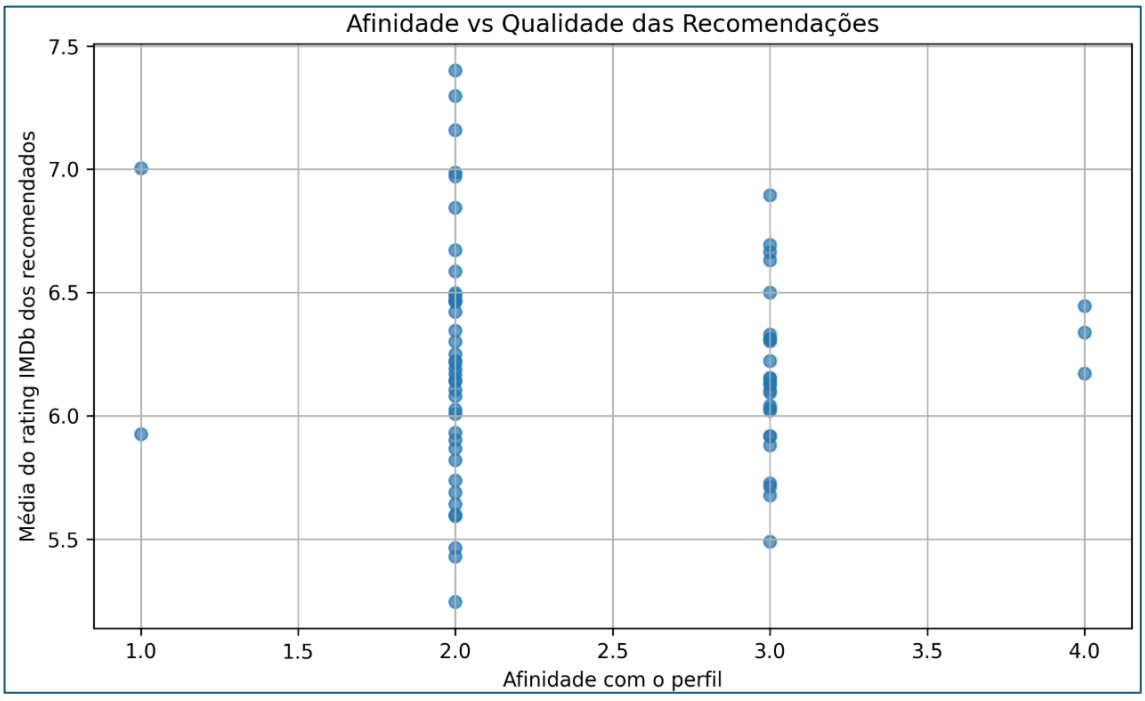
O modelo apresenta um desempenho relativamente adequado. Inicialmente nossa análise incluía apenas os gêneros dos filmes do usuário, porém na busca de melhorar nossas recomendações, inserimos na análise dos grupos maximais, os diretores e o elenco principal dos filmes. Essa mudança apresentou um aumento considerável no tempo de execução do algoritmo, que antes demorava cerca de segundos para definir os grupos, passou a demorar por volta de dois minutos na seleção.

Porém ainda sim apresenta um tempo adequado, que levando em consideração a quantidade de dados, permanece aceitável e viável.

Criamos também gráficos que transmitem de forma visual, informações sobre os resultados obtidos através das recomendações ao usuário.

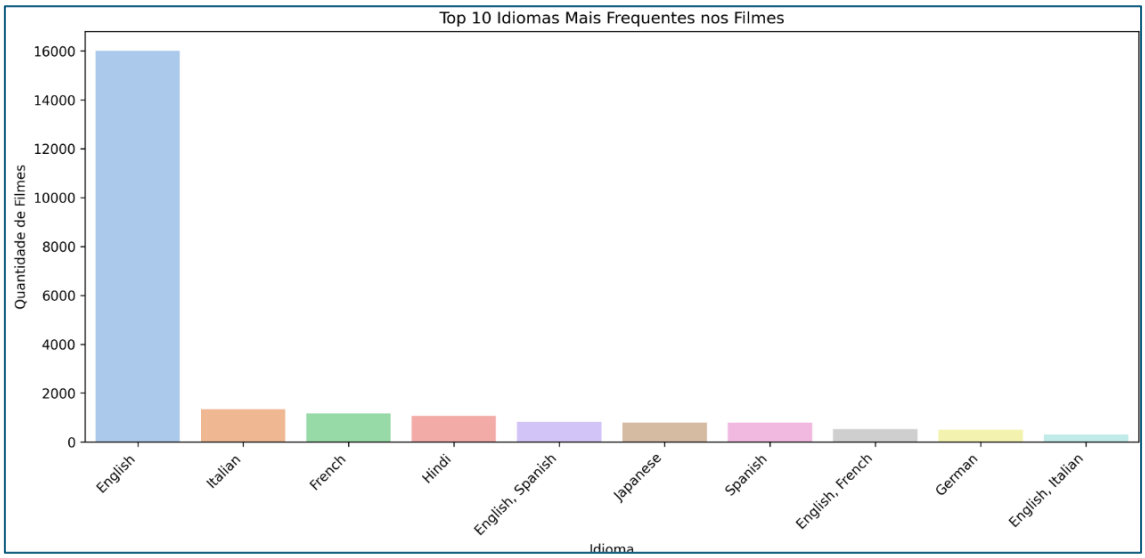


Nessa imagem temos os conjuntos de filmes por gênero, sendo que quanto maior a parcela do gênero maior a quantidade de filmes dessa categoria no conjunto final de recomendações.



Além disso temos esse gráfico que estabelece uma relação entre o valor de afinidade do conjunto de recomendação com a nota dos filmes presentes nesse conjunto. Uma análise possível é de que nem sempre um maior valor de afinidade quer dizer que as notas dos filmes do conjunto serão maiores, pois as recomendações dependem muito dos filmes

assistidos pelo usuário. No caso da imagem as maiores notas de filmes estão no conjunto com afinidade de valor 2.



Percebemos também como a base de dados utilizada é em sua grande maioria de filmes em inglês, logo as recomendações serão praticamente somente de filmes em inglês, o que limita o agrado dos usuários dependendo de suas preferências. Logo na busca de melhorias no nosso algoritmo e melhores recomendações, acreditamos que uma base de dados mais variada forneça resultados mais ecléticos.

6- Conclusão

Desenvolvemos diversos novos conhecimentos com esse trabalho, desde uma melhoria nas habilidades de manipulação de dados, necessária nas etapas de pré-processamento, até descobertas de algoritmos e modelos da área de Machine Learning.

Nossa maior dificuldade foi entender adequadamente o modelo utilizado na busca de introduzir novos dados na análise, além disso, tivemos alguns problemas na formatação das tabelas que variavam muito dependendo da recomendação e tamanho dos títulos e nomes. Esse problema foi superado utilizando bibliotecas de formatação que tornaram nossas tabelas mais harmônicas e agradáveis ao usuário, mesmo que visualizadas em linhas de comando.

Uma melhoria que desejamos realizar é a introdução de uma interface interativa para que o usuário comum possa interagir com nosso algoritmo, além de uma melhoria na análise de dados para levar em consideração ainda mais fatores na definição dos conjuntos.