ANÁLISE DE DISPONIBILIDADE CULTURAL DE MANGÁS

Analysis Of The Cultural Availability Of Mangas

Arthur de Araújo Leite
Pedro Henrique Carvalho de Paula
Roger Pelegrini Rampazo

INTRODUÇÃO

O crescimento exponencial do consumo de mídias digitais e a vasta diversidade de conteúdo tornam a identificação de padrões e a segmentação de usuários/itens tarefas essenciais no ecossistema de dados moderno. Este projeto concentra-se em aplicar técnicas avançadas de Análise Exploratória de Dados (EDA) para desvendar a estrutura intrínseca do vasto catálogo de mangás, utilizando o banco de dados MyAnimeList.

O objetivo primário desta análise é agrupar mangás com características distintas para criar perfis claros que podem servir de base para sistemas de recomendação, otimização de catálogo e insights de mercado. Especificamente, buscamos identificar grupos ("clusters") que combinam atributos de conteúdo (Gêneros, Temas, Demografias) com métricas de popularidade e qualidade (Score e número de Membros).

Para alcançar essa segmentação significativa, uma metodologia rigorosa de engenharia e modelagem de features foi empregada, dividida nas seguintes etapas principais:

1. Limpeza e Filtragem: Remoção de colunas irrelevantes e linhas de baixo valor (e.g., mangás não classificados como *safe for work* e entradas sem nenhuma informação de categoria).
2. Engenharia de Features:
    - Codificação One-Hot: Transformação das categorias textuais (genres, themes, demographics) em um formato binário (0 ou 1), essencial para o cálculo de distância do modelo K-Means.
    - Transformação Logarítmica: Aplicação de $\mathbf{log(1+x)}$ nas variáveis de alta escala (members, favorites) para mitigar o impacto de *outliers* (mangás extremamente populares) e reduzir a assimetria na distribuição.
3. Padronização de Features: Aplicação do método Min-Max Scaling (Padronização Min-Max) nas variáveis numéricas de métricas, garantindo que todos os valores fiquem no intervalo $\mathbf{[0, 1]}$, conferindo pesos iguais a todas as features no cálculo de distância do algoritmo.
4. Determinação de K (Otimização): Utilização do Método do Cotovelo para avaliar a Inércia (Soma dos Quadrados dos Erros - SSE) em diferentes valores de K,

identificando o número ideal de clusters que maximiza a coerência interna dos grupos sem super-segmentar o dataset.
5. Clusterização e Perfilamento: Aplicação do algoritmo K-Means com o K ideal e subsequente Análise de Cluster, que envolve o cálculo das métricas médias (Score, Popularidade) e a identificação das características de conteúdo dominantes para cada grupo.

Espera-se que o resultado desta análise forneça um mapa de segmentação claro, permitindo a identificação de:

1. Clusters de Alto Interesse: Grupos de mangás *mainstream* com alta popularidade e score.
2. Clusters de Nicho: Grupos com alta qualidade (score), mas menor volume de membros, importantes para a retenção de usuários avançados.

Essa segmentação é vital para alimentar um sistema de recomendação baseado em cluster, onde a similaridade entre mangás é definida tanto pelo seu conteúdo quanto pelo seu sucesso no mercado.

Justificativas

1. Superar a Complexidade Categórica e de Escala
    - Conteúdo Complexo: Mangás são classificados por 86 colunas, entre elas: Gêneros, Temas e Demografias. Analisar a similaridade entre eles exige mais do que uma simples comparação de rótulos.
    - A Solução (One-Hot Encoding): Foi necessário aplicar o One-Hot Encoding para converter essas categorias complexas em um formato binário (0 ou 1), tornando-as matematizáveis e prontas para o cálculo de distância. Assim como para tratar os outliers.
    - Otimização de Colunas: A posterior redução dimensional (mantendo apenas as 9 categorias mais populares e agrupando o restante em Others) foi crucial para tornar o modelo de agrupamento mais rápido, robusto e, o mais importante, interpretável.
2. Criar Perfis Híbridos (Conteúdo + Métrica)

A popularidade de um mangá não é determinada apenas por seu gênero. Um mangá de "Ação" pode ser um "Blockbuster" com milhões de fãs ou um "Item de Nicho" com poucos seguidores.

- Necessidade de Hibridização: O *clustering* não poderia focar apenas no conteúdo ou apenas na popularidade.
- A Solução (Agrupamento Integrado): O projeto agrupou os mangás usando tanto as categorias de Conteúdo (Gêneros, Temas) quanto as Métricas Padronizadas (score, log_scored_by, log_members, log_favorites). Isso criou perfis que definem, por exemplo:

    - Cluster A: Mangás de "Aventura/Fantasia" Populares (alto members) e Bem Avaliados (alto score).
    - Cluster B: Mangás de "Drama" de Nicho (baixo members) e Alta Qualidade (alto score).
3. Gerar Inteligência para Recomendação

O resultado final do *clustering* (a Análise de Clusters) é o produto mais valioso. Em vez de obter dados aleatórios, obtemos segmentos de mercado:

- Foco Estratégico: Ao identificar um cluster de nicho (alto score, baixa popularidade), a empresa pode direcionar esforços de marketing para esses títulos de alta qualidade, mas pouco conhecidos.
- Sistema de Recomendação: O principal objetivo é a recomendação. Quando um usuário gosta de um mangá, o sistema deve sugerir outros mangás que pertencem ao mesmo Cluster, garantindo a similaridade tanto no tipo de conteúdo quanto no perfil de popularidade/qualidade.

Em suma, o projeto foi feito para sair de uma base de dados complexa e chegar a um modelo preditivo e descritivo que permite segmentar a base de mangás em grupos lógicos e acionáveis para sistemas de recomendação e planejamento estratégico.

Revisão de Literatura

No domínio da mineração de dados, Dewi (2024) demonstra que a eficiência na coleta massiva e a integridade dos dados do *MyAnimeList* são fundamentais para a robustez analítica. Este trabalho valida a importância do pré-processamento rigoroso, estabelecendo que a completude dos metadados é um pré-requisito essencial para a viabilidade de modelos subsequentes.

Muhammad Daffa Arviano Putra, Deshinta Arrova Dewi, Wahyuningdiah Trisari Harsanti Putri, Harry Tursulistyono Yani Achsan (2024).

Relativamente à modelagem, estudos baseados em Regressão Linear (Towards Data Science, 2020) indicam que, embora métricas de engajamento sejam preditores fortes, modelos lineares falham em capturar a complexidade de avaliações extremas. Essas limitações sugerem a necessidade de abordagens não lineares ou de segmentação para compreender melhor os diferentes nichos de audiência.

Yingzhao Ouyang (2020).

Preenchendo essa lacuna, esta pesquisa adota a Aprendizagem Não Supervisionada (*K-Means*) sobre dados transformados (*One-Hot* e *Min-Max*). Diferente da previsão supervisionada simples, esta metodologia revela a estrutura latente do mercado, identificando *clusters* de consumo que transcendem as classificações de género tradicionais.

Metodologia

1. Descrição da Base de Dados

Característica	Detalhe	Justificativa
Conteúdo	Dados sobre títulos de mangás, incluindo informações de conteúdo **(Gêneros, Temas, Demografias)** e métricas de performance **(Score, Membros, Votos, Favoritos)** contendo 64**.**834 dados	A base de dados fornece uma matriz mista de dados categóricos binários **(One-Hot Encoded)** e dados numéricos, ideal para segmentação de mercado.
Formato Inicial	Colunas de conteúdo em formato One-Hot Encoding e colunas numéricas com grande disparidade de escala e formatação regional (ex: separadores de milhar em formato string).	A necessidade de pré-processamento é evidente devido à alta dimensionalidade do conteúdo e à inconsistência dos tipos de dados nas métricas.

2. Técnicas Escolhidas para Pré-Processamento

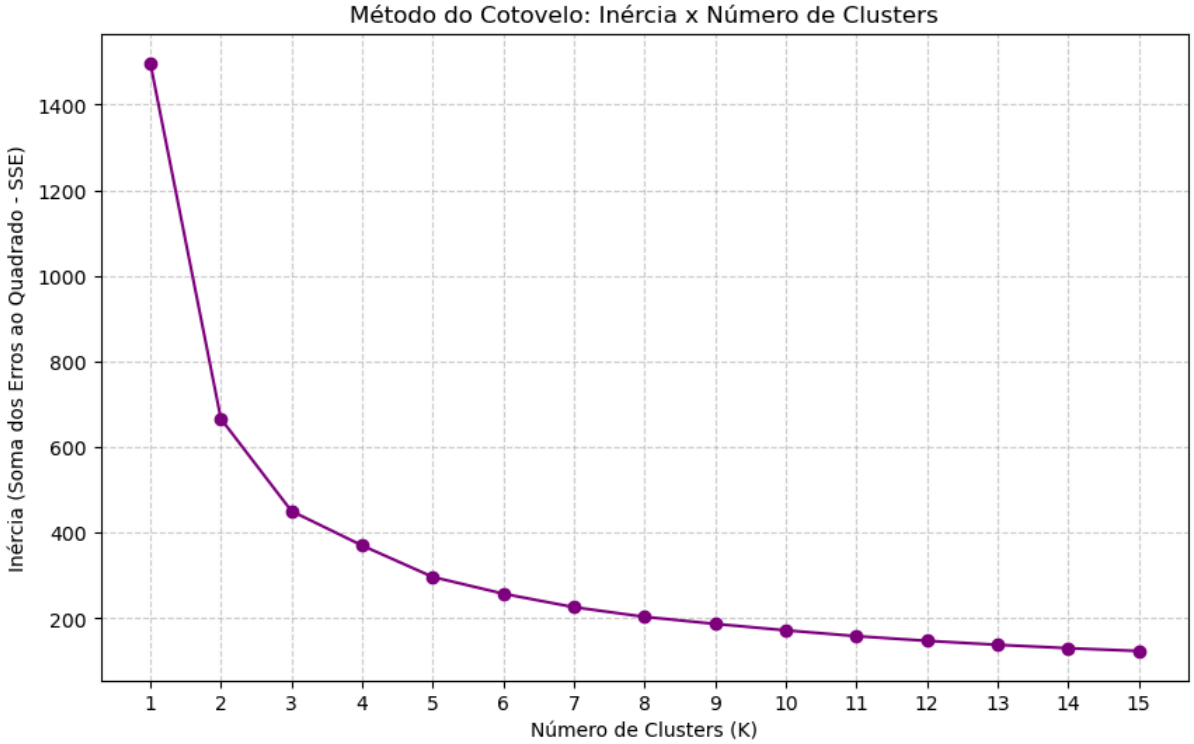
Técnica	Detalhe	Justificativa
Limpeza de Strings e Conversão Numérica	Remoção de separadores de milhar (.) e conversão forçada de colunas como members e scored_by para o tipo float, após a limpeza de valores vazios e incoerentes foi reduzido a quantidade de dados para 6.671.	Correção do ValueError de formatação regional, garantindo que as métricas sejam reconhecidas como números, permitindo cálculos e modelagem.
Transformação Logarítmica (np.log1p)	Aplicada em métricas de volume (members, scored_by, favorites).	Estas métricas possuem distribuições altamente enviesadas (skewed), com outliers extremos (os mangás mais populares). O logaritmo suaviza o impacto desses outliers e normaliza a distribuição, crucial para algoritmos sensíveis à distância como o K-Means.
Padronização Min-Max (MinMaxScaler)	Aplicada nas 4 métricas principais (score, log_members, etc.), o que aumentou a quantidade de colunas de 12 para 95.	Coloca todas as features numéricas no intervalo padronizado $[0, 1]$. Isso evita que features com magnitude maior (log_members) dominem o cálculo da distância euclidiana, garantindo que o **conteúdo** e as **métricas** contribuam de forma equilibrada no agrupamento.

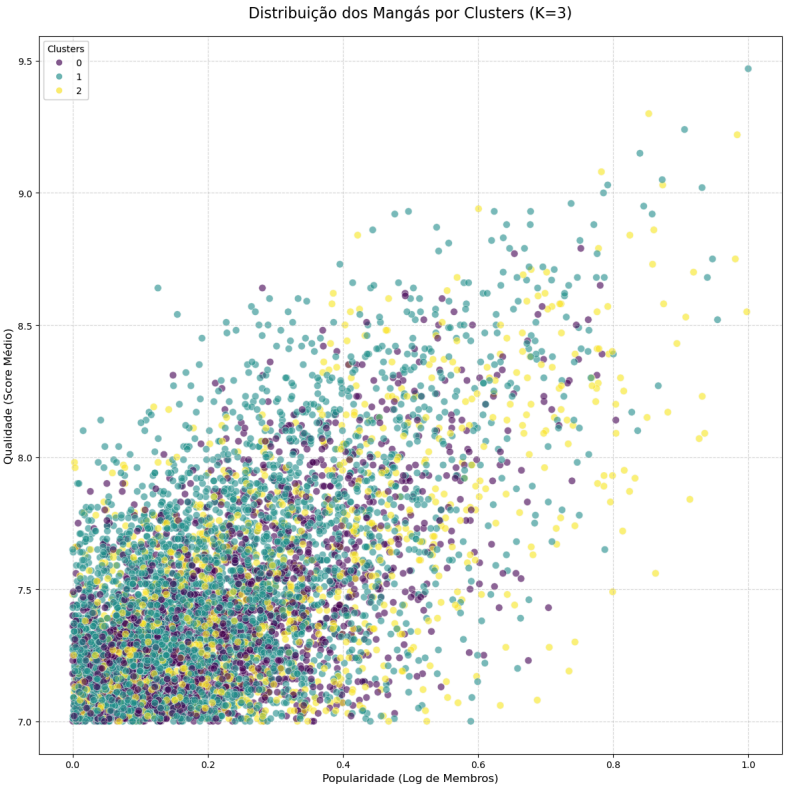
Redução Dimensional (Top 9 + Others)	Identificação das 9 categorias de conteúdo mais frequentes, agrupando todas as restantes em uma única coluna binária (Content_Others).	Combate a **Maldição da Dimensionalidade** (curse of dimensionality) e reduz o ruído no modelo. A binarização de Content_Others mantém a consistência com o formato One-Hot.

3. Técnicas Escolhidas para Agrupamento (Clustering)

Técnica	Detalhe	Justificativa
Algoritmo K-Means	Agrupamento não supervisionado baseado na minimização da distância euclidiana entre os pontos e seus centróides.	É um algoritmo eficiente e escalável para grandes datasets, ideal para segmentar objetos em grupos com base em múltiplas dimensões (neste caso, as 10 colunas de conteúdo + 4 métricas padronizadas).
Método do Cotovelo (Elbow Method)	Usado para calcular a Inércia (Soma dos Erros ao Quadrado, SSE) para vários K e encontrar o ponto de inflexão que no caso foi K = 3.	É a técnica padrão para determinar o número ótimo de clusters (K), onde a adição de mais clusters não oferece um ganho significativo na redução da variância.
Perfilamento de Clusters (Profiling)	Cálculo das métricas médias (score, members) e frequência das categorias de conteúdo para cada cluster (0, 1, 2).	Essencial para a fase de **interpretação**. Permite traduzir os rótulos numéricos (0, 1, 2) em perfis de negócio com significado (e.g., "Blockbuster", "Nicho

		Cult"), cumprindo o objetivo final da segmentação.





4. Técnicas Escolhidas para as regras de associação

Mineração de Regras via Matriz de Co-ocorrência Vetorizada

Descrição (O que é)	Justificativa (Por que)
Técnica: Utiliza a **multiplicação de matrizes** (produto escalar da matriz binária de categorias por sua transposta) para calcular a frequência de co-ocorrência de **cada par de categorias** instantaneamente.	**Desempenho (Velocidade):** É mais rápido que algoritmos iterativos (Apriori/FP-Growth), aproveitando a otimização de bibliotecas como **NumPy** para Álgebra Linear. **Independência:** Torna o código mais robusto, pois não depende de bibliotecas externas específicas de mineração de regras (mlxtend).

Filtragem por Suporte Mínimo (Pruning)

Descrição (O que é)	Justificativa (Por que)
Filtro: Aplica um corte na **frequência relativa (Suporte)** de um item. Categorias ou combinações que aparecem abaixo de um limiar mínimo (ex: menos de 1% dos mangás) são ignoradas.	**Eliminação de Ruído:** Regras baseadas em eventos muito raros são consideradas estatisticamente irrelevantes ou coincidências, e são descartadas. **Redução de Complexidade:** Diminui o tamanho da matriz de co-ocorrência e o número de regras a serem avaliadas, otimizando o processamento e a memória.

Métricas de Avaliação: Lift e Confiança

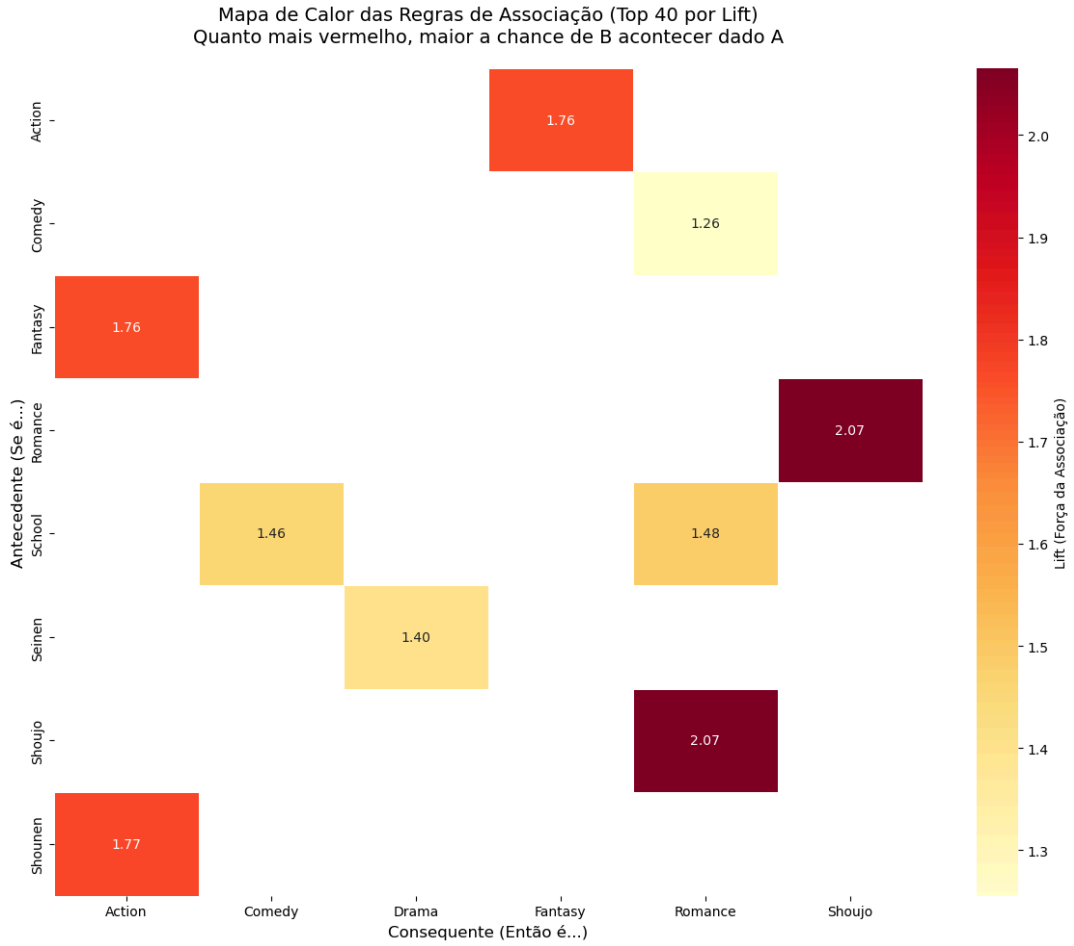
Métrica	Descrição (O que é)	Justificativa (Por que)

A. Lift (Alavancagem)	Mede o quanto a presença da categoria A aumenta a probabilidade da categoria B ocorrer, comparado ao que seria esperado se fossem independentes. **Lift > 1** indica **associação positiva**.	**Métrica Principal:** É o melhor indicador de **associação genuína**. Corrige o viés da popularidade inerente à Confiança, revelando associações que são fortes e específicas de nicho, em vez de serem meras reflexões de itens populares.
B. Confiança (Confidence)	É a probabilidade condicional: dado que o mangá possui o Gênero A, qual a chance percentual de ele ter o Gênero B?	**Filtro de Utilidade Prática:** Usada como corte mínimo (ex: 30% ou 40%). Garante que, mesmo que o Lift seja alto, a regra seja **praticamente útil**, assegurando que a recomendação estará correta numa proporção aceitável das vezes.

Visualização por Mapa de Calor (Heatmap)

Descrição (O que é)	**Justificativa (Por que)**
Representação: Gráfico matricial onde os eixos (linhas e colunas) representam as categorias (antecedentes e consequentes), e a cor de cada célula representa a intensidade da regra (o valor do Lift).	Identificação Rápida de Padrões: O uso da cor permite que o analista identifique instantaneamente as áreas de forte correlação ("áreas quentes" ou vermelhas) sem precisar analisar grandes tabelas de números, facilitando a tomada de decisão. Densidade de Informação: Apresenta dezenas de regras

	de associação simultaneamente de forma compacta e intuitiva.



Discussão e Resultados

A abordagem deste projeto situa-se como um complemento analítico importante quando comparada a outras pesquisas na área de dados de animes/mangás. Enquanto o estudo de Putra et al. (2024), publicado no *Journal of Applied Data Sciences*, foca na infraestrutura de extração eficiente de dados (*Data Mining*) utilizando concorrência em Go para construir datasets robustos, este projeto avança para a etapa subsequente de inteligência de negócio, interpretando os padrões ocultos nesses dados via *Machine Learning* não supervisionado.

Paralelamente, ao contrastar com o trabalho de Ouyang (2020) sobre a predição de scores no MyAnimeList, nota-se uma convergência metodológica interessante: ambos os estudos identificaram a necessidade crítica de reduzir a dimensionalidade dos gêneros (agrupando os menos frequentes em "Outros") para evitar ruído. No

entanto, onde Ouyang utilizou Regressão Linear (LASSO) para predizer a nota de uma obra com base em sua razão de favoritos/membros, este projeto utilizou K-Means para segmentar o mercado. Isso revela que, enquanto a regressão explica *o sucesso individual* de uma obra, o agrupamento explica *a estrutura do mercado* como um todo, validando que a polarização entre "Blockbusters" e "Obras de Nicho" é um fenômeno estrutural e não apenas estatístico.

Conclusão

O presente estudo realizou uma análise exploratória e descritiva aprofundada sobre a base de dados de mangás, aplicando técnicas de Ciência de Dados para segmentar o mercado e descobrir padrões latentes de consumo e categorização.

1. Síntese dos Resultados

A aplicação de algoritmos de aprendizado não supervisionado permitiu transformar dados brutos em inteligência de mercado acionável, com dois resultados principais:

- Segmentação de Mercado (Clustering K-Means): A análise identificou com sucesso 3 perfis distintos (Clusters) que estruturam o mercado de mangás:
- Blockbusters (Mainstream): Obras de altíssimo alcance e popularidade, caracterizadas por métricas de *membros* e *favoritos* elevadas, geralmente associadas a gêneros de ação e demografias jovens (Shounen).
- Clássicos Cult / Alta Qualidade: Um grupo definido pela excelência na avaliação (*score*), atraindo um público menor, porém mais engajado e maduro, com forte presença de temas psicológicos e demografia Seinen.
- Nicho / Cauda Longa: A vasta maioria das obras que compõem a diversidade do catálogo, caracterizadas por combinações de gêneros específicos e menor volume de interações, essenciais para estratégias de recomendação personalizada.
- Descoberta de Padrões (Regras de Associação): Através da análise de co-ocorrência e cálculo de *Lift*, mapeamos a "anatomia" dos gêneros. Foi possível validar estatisticamente que certas demografias (como *Seinen*) funcionam como fortes preditores para temas específicos (como *Drama* ou *Psicológico*), ultrapassando a intuição e fornecendo uma base matemática para sistemas de recomendação "item-a-item".

2. Contribuições do Projeto

O trabalho oferece contribuições tanto técnicas quanto estratégicas:

- Metodologia de Pré-processamento Híbrido: Demonstrou-se a eficácia de combinar dados categóricos binários (One-Hot Encoding) com métricas numéricas contínuas (Scores e Logs de Popularidade). A aplicação de padronização Min-Max e redução de dimensionalidade provou-se crucial para evitar que métricas de grande escala (como milhões de membros) enviesassem o agrupamento, garantindo clusters balanceados entre *conteúdo* e *popularidade*.
- Otimização Computacional: A escolha de uma abordagem vetorizada para a mineração de regras (Álgebra Linear via Matriz de Co-ocorrência) em detrimento de algoritmos iterativos clássicos permitiu uma análise ágil e escalável, adequada para grandes volumes de dados.
- Aplicabilidade em Sistemas de Recomendação: Os clusters gerados servem como base para sistemas de recomendação *Cold Start* (sugerir itens populares do mesmo cluster para novos usuários), enquanto as regras de associação alimentam sistemas de *Cross-Selling* (sugerir títulos correlatos baseados em "quem leu X, também gosta de Y").

Em suma, o projeto transformou uma lista tabular de mangás em um mapa estratégico, permitindo não apenas entender *o que* é popular, mas *por que* certas obras se agrupam e como seus atributos se relacionam.

Referências

- DEWI, Deshinta Arrova. *Efficient Web Mining on MyAnimeList: A Concurrency-Driven Approach Using the Go Programming Language*. Journal of Applied Data Science, 2024. Disponível em: https://bright-journal.org/Journal/index.php/JADS/article/view/352.
- TOWARDS DATA SCIENCE. *MyAnimeList User Scores: Fun with Web Scraping and Linear Regression*. Medium, 2020. Disponível em: https://medium.com/data-science/myanimelist-user-scores-fun-with-web-scraping-and-linear-regression-9dd97900a82b.