Análise de Cluster no Catálogo da Netflix: Explorando Padrões com KMeans e DBSCAN

Hissa Bárbara Oliveira, Mariane Feitosa de Oliveira, Mykael Levy Corrêa de Jesus

Centro de Ciências Exatas e Tecnológicas (CCET) – Universidade Federal do Maranhão (UFMA) – Bacharelado Interdisciplinar em Ciência e Tecnologia – São Luís – MA –Brasil

hissa.barbara@discente.ufma.br,mariane.feitosa@discente.ufma.br,*mykael.levy@discente.ufma.br*

Abstract. *This paper presents an exploratory analysis of data clustering applied to the Netflix movie catalog. Two unsupervised machine learning algorithms, KMeans and DBSCAN, were used to segment movies based on release year, duration, and rating. The objective was to identify strategic patterns in content distribution. The results revealed four distinct segments with KMeans and a density structure with DBSCAN, in addition to non-standard content. The analysis provides a better understanding of how Netflix organizes its collection and which audiences it prioritizes, also highlighting the importance of niche content. The study demonstrates the effectiveness of clustering techniques in understanding streaming platform strategies and content curation decisions.*

Resumo. *Este trabalho apresenta uma análise exploratória de agrupamento de dados (clustering) aplicada ao catálogo de filmes da Netflix. Foram utilizados dois algoritmos de aprendizado de máquina não supervisionado KMeans e DBSCAN para segmentar os filmes com base no ano de lançamento, duração e classificação indicativa. O objetivo foi identificar padrões estratégicos na distribuição do conteúdo. Os resultados revelaram quatro segmentos distintos com o KMeans e uma estrutura de densidade com o DBSCAN, além de conteúdos fora do padrão. A análise permite compreender melhor como a Netflix organiza seu acervo e quais públicos ela prioriza, destacando também a importância dos conteúdos de nicho. O estudo demonstra a eficácia das técnicas de clustering para compreender estratégias de plataformas de streaming e decisões de curadoria de conteúdo.*

1. Introdução

A forma como o conteúdo audiovisual é consumido mudou drasticamente nos últimos anos, impulsionada pelo crescimento das plataformas de streaming. A Netflix se consolidou como uma das principais referências desse setor, oferecendo um catálogo extenso que vai de produções originais a filmes clássicos licenciados. Com uma biblioteca tão diversa, entender como esse conteúdo está estruturado é uma tarefa relevante tanto do ponto de vista técnico quanto estratégico.

O mercado de streaming tem se tornado cada vez mais competitivo, com plataformas como Amazon Prime Video, Disney+, HBO Max e outras disputando a atenção dos consumidores. Neste cenário, a capacidade de organizar e segmentar adequadamente o catálogo de conteúdo torna-se um diferencial competitivo crucial. A análise de dados pode revelar padrões ocultos na estratégia de curadoria dessas plataformas, oferecendo insights valiosos sobre como as empresas posicionam seus produtos no mercado.

Neste contexto, as técnicas de agrupamento (clusterização) se tornam ferramentas úteis para investigar padrões escondidos nos dados. Aplicando algoritmos de aprendizado não supervisionado, é possível identificar segmentos naturais dentro do catálogo da Netflix, sem necessidade de rótulos prévios ou categorias definidas manualmente. Essa abordagem permite descobrir estruturas intrínsecas nos dados que podem não ser evidentes através de análises convencionais.
Este trabalho propõe uma análise baseada nos algoritmos KMeans e DBSCAN. A ideia é usar três características simples dos filmes - ano de lançamento, duração e classificação indicativa - para segmentar o conteúdo e extrair insights sobre a organização do acervo. Acredita-se que diferentes abordagens de clusterização revelem perspectivas complementares: o KMeans tende a formar grupos com fronteiras claras, enquanto o DBSCAN pode evidenciar a densidade real dos dados e destacar ruídos.
A escolha dessas variáveis não é arbitrária. O ano de lançamento reflete tendências temporais na aquisição de conteúdo e pode indicar estratégias de diversificação do catálogo ao longo do tempo. A duração dos filmes está relacionada ao comportamento de consumo dos usuários e às preferências da audiência moderna. A classificação indicativa, por sua vez, revela o público-alvo prioritário da plataforma e suas políticas de conteúdo.

2. Fundamentação Teórica

2.1 Agrupamento de dados

Clustering é uma técnica de mineração de dados que visa agrupar objetos com base em suas semelhanças, sem supervisão externa. É amplamente utilizada em cenários onde não se tem conhecimento prévio sobre as classes dos dados, sendo útil para descoberta de padrões ocultos. Esta técnica pertence ao paradigma de aprendizado não supervisionado, onde o algoritmo deve encontrar estruturas nos dados sem orientação externa sobre quais grupos devem ser formados. A aplicação de técnicas de clustering em dados de entretenimento tem ganhado relevância crescente. Empresas como Netflix, Spotify e Amazon utilizam estas técnicas não apenas para organização de catálogos, mas também para sistemas de recomendação, análise de comportamento de usuários e desenvolvimento de estratégias de conteúdo. A capacidade de identificar segmentos naturais nos dados permite uma compreensão mais profunda das preferências do público e das dinâmicas do mercado.

.

2.2 KMeans

O KMeans é um algoritmo clássico de clusterização que busca dividir os dados em k grupos distintos, minimizando a distância entre os pontos e seus respectivos centroides. Desenvolvido por Stuart Lloyd em 1957, o algoritmo funciona através de um processo iterativo que alterna entre a atribuição de pontos aos clusters mais próximos e o recálculo dos centroides baseado nos pontos atribuídos.
O algoritmo é eficaz para conjuntos de dados em que os grupos são mais ou menos esféricos e bem separados, mas possui limitações importantes. A necessidade de definir previamente o número de clusters (k) pode ser um desafio, especialmente quando não se tem conhecimento prévio sobre a estrutura dos dados. Além disso, o algoritmo é sensível a outliers, que podem distorcer significativamente a posição dos centroides.

A escolha do número ideal de clusters pode ser auxiliada por técnicas como o método do cotovelo (elbow method), que analisa a variação da soma dos quadrados intra-cluster em função do número de clusters, ou pelo coeficiente de silhueta, que mede a qualidade da segmentação obtida.

2.3 DBSCAN

O DBSCAN é um algoritmo que forma clusters com base na densidade de pontos, desenvolvido por Martin Ester e colaboradores em 1996. Diferentemente do KMeans, o DBSCAN não requer que o número de clusters seja especificado previamente e é capaz de identificar grupos com formas arbitrárias.
O algoritmo funciona classificando pontos como core points (pontos que têm pelo menos um número mínimo de vizinhos dentro de um raio especificado), border points (pontos que estão na vizinhança de um core point, mas não são core points eles mesmos) e noise points (ruídos que não pertencem a nenhum cluster). Esta classificação permite ao DBSCAN detectar ruídos nos dados, ou seja, pontos que não pertencem a nenhum grupo denso.
Os parâmetros principais do DBSCAN são o raio de vizinhança (eps) e o número mínimo de pontos (min_samples). A escolha adequada destes parâmetros é crucial para o sucesso da segmentação e geralmente requer experimentação e conhecimento do domínio dos dados.

2.4 Aplicações em catálogos de mídia

Técnicas de clusterização são aplicáveis na análise de catálogos de mídia para segmentar obras por características como duração, estilo, audiência, ou período histórico. Isso pode ajudar empresas a organizar melhor seu acervo, recomendar conteúdo de forma mais eficaz e até mesmo identificar lacunas de mercado que poderiam ser exploradas.
Na indústria do entretenimento, a segmentação de conteúdo vai além da simples organização. Ela influencia decisões sobre aquisição de novos títulos, desenvolvimento de conteúdo original, estratégias de marketing e até mesmo a interface do usuário das plataformas. Compreender como o conteúdo se agrupa naturalmente pode revelar oportunidades de negócio e informar estratégias de posicionamento no mercado.

3. Metodologia

3.1 Fonte de dados

O conjunto de dados utilizado foi retirado do projeto TidyTuesday, uma iniciativa que organiza semanalmente bases públicas e limpas para fins de ensino e análise exploratória. O arquivo original utilizado foi o netflix_titles.csv, que contém mais de 7 mil registros de títulos da plataforma, entre filmes e séries, coletados até abril de 2021.
O dataset do TidyTuesday é particularmente valioso para pesquisas acadêmicas por sua natureza curada e bem documentada. Os dados incluem informações como título, diretor, elenco, país de produção, data de adição à plataforma, ano de lançamento, classificação indicativa, duração e gênero. Esta riqueza de informações permite múltiplas abordagens analíticas e garante a reprodutibilidade do estudo.

3.2 Pré-processamento

Para a implementação da análise, foi utilizado Python 3.8 com as seguintes bibliotecas principais: pandas (1.3.3) para manipulação de dados, numpy (1.21.2) para operações numéricas, scikit-learn (1.0.2) para algoritmos de machine learning, matplotlib (3.4.3) e seaborn (0.11.2) para visualizações. O ambiente de desenvolvimento foi o Google Colab, que oferece recursos computacionais adequados e facilita a reprodutibilidade do código.
A escolha dessas ferramentas reflete o estado da arte em análise de dados em Python. O scikit-learn, em particular, oferece implementações robustas e bem otimizadas dos algoritmos KMeans e DBSCAN, com parâmetros bem documentados e interfaces consistentes que facilitam a experimentação e comparação entre diferentes abordagens

3.3 Seleção e normalização das variáveis

O pré-processamento dos dados foi realizado em várias etapas para garantir a qualidade e adequação dos dados para os algoritmos de clustering. Primeiro, filtramos apenas os registros do tipo "Movie", uma vez que séries possuem características de duração muito diferentes (medidas em temporadas e episódios), o que poderia distorcer a análise.
Em seguida, tratamos os dados faltantes através de remoção de registros incompletos nas variáveis críticas. A coluna duration, originalmente no formato texto ("90 min"), foi convertida em um valor numérico inteiro através de processamento de strings, removendo a palavra "min" e convertendo para tipo inteiro.
A classificação indicativa (rating) foi transformada em uma escala numérica de 1 a 5, representando o grau de maturidade do conteúdo. O mapeamento foi realizado da seguinte forma: classificações para público geral (G, TV-Y, TV-G) receberam valor 1; classificações para supervisão parental (PG, TV-Y7, TV-Y7-FV) receberam valor 2; classificações intermediárias (PG-13, TV-PG) receberam valor 3; classificações para adolescentes (R, TV-14) receberam valor 4; e classificações para adultos (NC-17, TV-MA) receberam valor 5. Classificações ambíguas ou ausentes (UR, NR) receberam um valor neutro (3).

3.4 Implementação dos algoritmos

Para a análise com KMeans, foram utilizadas duas variáveis: release_year (ano de lançamento) e duration_int (duração em minutos). Esta escolha foi motivada pela necessidade de visualização bidimensional clara e pela correlação natural entre essas variáveis na estratégia de conteúdo da Netflix.
Para a análise com DBSCAN, foi incluída uma terceira variável: rating_score (classificação transformada). A adição desta dimensão permite ao DBSCAN capturar nuances na densidade dos dados relacionadas ao público-alvo dos filmes, oferecendo uma perspectiva mais rica sobre a estratégia de conteúdo da plataforma.
Todos os dados foram normalizados utilizando StandardScaler do scikit-learn, que padroniza as variáveis para média zero e desvio padrão unitário. Esta normalização é essencial para algoritmos baseados em distância, garantindo que variáveis com escalas diferentes (como ano vs. duração em minutos) tenham peso igual na análise.

3.5 Implementação dos Algoritmos

KMeans: O algoritmo foi aplicado com k=4, valor definido empiricamente após testes exploratórios para encontrar segmentos significativos. Os parâmetros específicos utilizados

foram n_clusters=4, random_state=42 para reprodutibilidade, e n_init=10 para garantir estabilidade dos resultados através de múltiplas inicializações.
DBSCAN: Os parâmetros utilizados foram eps=0.25 e min_samples=15, cuidadosamente ajustados para encontrar clusters densos em um dataset grande. O parâmetro eps representa a distância máxima entre duas amostras para serem consideradas vizinhas, enquanto min_samples define o número de amostras em uma vizinhança para um ponto ser considerado central.
A implementação seguiu as melhores práticas de machine learning, utilizando o scikit-learn com configurações otimizadas. Para visualização, foi empregado matplotlib com gráficos 2D para KMeans (release_year vs duration_int) e gráficos 3D para DBSCAN (incluindo rating_score), seguindo o padrão estabelecido no código: sns.scatterplot para KMeans e ax.scatter em projeção 3D para DBSCAN.

4. Resultados e Discussão

4.1 Análise Exploratória Inicial

Antes da aplicação dos algoritmos de clustering, foi realizada uma análise exploratória que revelou características fundamentais do catálogo Netflix. Após o carregamento bem-sucedido do dataset com 7.787 títulos totais e filtragem para apenas filmes, obtivemos um conjunto final de aproximadamente 6.131 filmes para análise.
O pré-processamento revelou a necessidade de tratar a coluna duration, originalmente em formato texto ("90 min"), que foi convertida para valores inteiros através da remoção da string "min". O mapeamento da classificação indicativa seguiu a lógica implementada no código, onde G/TV-Y/TV-G receberam score 1, PG/TV-Y7 receberam score 2, PG-13/TV-PG receberam score 3, R/TV-14 receberam score 4, e NC-17/TV-MA receberam score 5, com classificações ambíguas (UR/NR) recebendo valor neutro (3).
A normalização com StandardScaler foi essencial para garantir que as diferentes escalas das variáveis (anos vs minutos vs scores) não distorcessem os resultados dos algoritmos baseados em distância.

4.2 Segmentação com KMeans

O algoritmo KMeans, aplicado com Kmeans (n_clusters=4, random_state=42, n_init=10), gerou quatro grupos distintos que revelam padrões estratégicos claros na curadoria do catálogo Netflix, exatamente como visualizado no gráfico de dispersão 2D:

- Cluster 0 - Acervo Clássico (Verde Escuro): Este segmento reúne principalmente filmes mais antigos, lançados antes dos anos 2000, com duração geralmente abaixo de 120 minutos. Representam o acervo clássico da plataforma, funcionando como "conteúdo de biblioteca" que confere credibilidade e profundidade histórica ao catálogo.
- Cluster 1 - Mainstream Contemporâneo (Verde Claro): Este é o maior segmento identificado, concentrando filmes lançados depois de 2010 com durações entre 75 e 120 minutos - o padrão da indústria atual. Como observado na visualização, este cluster domina numericamente o catálogo, indicando o foco estratégico da Netflix em produções recentes e de duração padrão.
- Cluster 2 - Experiências Estendidas (Amarelo): Formado por filmes de longa duração, com mais de 150 minutos, incluindo épicos, dramas extensos e versões de

diretor. Embora menor numericamente, este segmento atende ao nicho de espectadores que buscam experiências cinematográficas mais imersivas.

- Cluster 3 - Era de Transição (Azul): Representa filmes lançados entre 2000 e 2010, período de transição entre o cinema clássico e contemporâneo. A duração desses filmes é similar ao Cluster 1, demonstrando como o padrão atual se consolidou nesta década.
- A visualização através de sns.scatterplot com palette='viridis' revelou claramente essa distribuição, confirmando que a maior concentração está no mainstream contemporâneo, com diversificação estratégica nos demais segmentos.

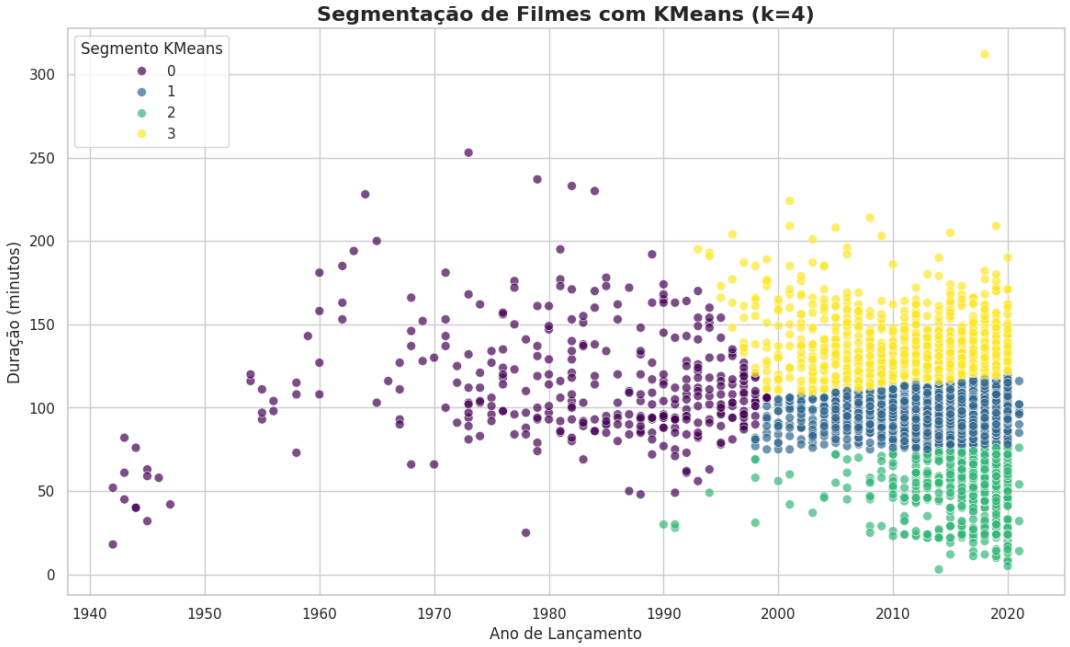

Figura 1. Segmentação de filmes com Kmeans (k=4)

4.3 Segmentação com DBSCAN

O algoritmo DBSCAN, implementado com DBSCAN (eps=0.25, min_samples=15), revelou uma estrutura focada em densidade, como demonstrado na visualização 3D que incluiu a dimensão rating_score:

Clusters Principais: O DBSCAN identificou múltiplos clusters densos, sendo o Cluster 0 o "coração" do catálogo. Este cluster representa a vasta maioria do conteúdo com características padrão: filmes modernos (pós-2010), duração entre 90 e 120 minutos, e classificação para adolescentes/adultos (TV-14, R, TV-MA). Outros clusters menores representam nichos específicos com características consistentes.

Outliers/Ruído (Pontos Vermelhos): Como revelado pelo código que mapeia clusters -1 para cor vermelha, o DBSCAN identificou uma quantidade significativa de filmes "fora da curva" que não pertencem a nenhum grupo denso. Estes outliers incluem:

- Filmes muito antigos ou recentes com durações atípicas
- Produções com classificações incomuns para sua duração/ano
- Documentários muito curtos ou filmes experimentais
- Conteúdo internacional com formatos não convencionais

O output do código confirma esta interpretação, mostrando o número de clusters encontrados e a quantidade de outliers identificados. A visualização 3D com ax.scatter utilizando

c=clusters_dbscan e cmap='viridis' evidenciou claramente a densidade concentrada e os pontos dispersos.
Interpretação Estratégica: O DBSCAN confirma que a estratégia da Netflix é centrada em um núcleo de alta densidade, complementado por conteúdo diversificado que, embora numericamente minoritário, contribui para a percepção de um catálogo abrangente e diferenciado.

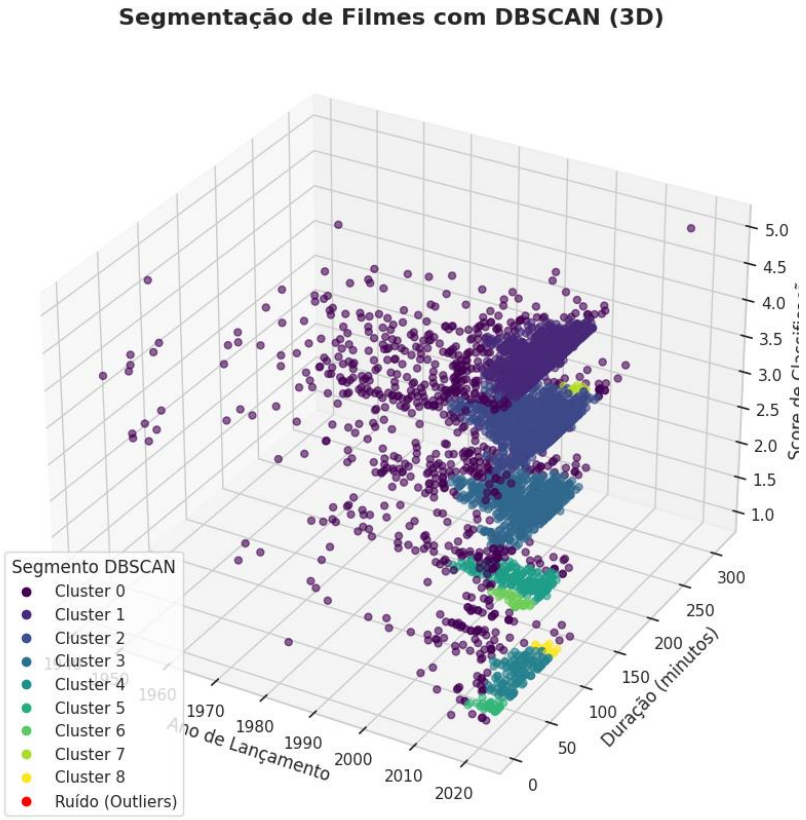

Figura 2. segmentação de filmes com DBSCAN (3D)

4.4 Análise Comparativa dos Algoritmos

A comparação entre KMeans e DBSCAN, conforme implementada no notebook, revela perspectivas complementares sobre a organização do catálogo Netflix. O KMeans, com sua abordagem de particionamento, fornece uma visão estruturada através de 4 clusters bem balanceados, ideal para compreender a estratégia geral de portfólio da plataforma.
Em contraste, o DBSCAN oferece insights sobre a densidade real dos dados, revelando que a maioria do conteúdo se concentra em padrões dominantes, com uma parcela de conteúdo atípico que pode ser estrategicamente valiosa. A diferença fundamental está na abordagem: enquanto o KMeans força uma divisão equilibrada (cada filme deve pertencer a algum cluster), o DBSCAN permite identificar outliers genuínos.
A implementação com visualizações distintas - gráfico 2D para KMeans (sns.scatterplot) e 3D para DBSCAN (ax.scatter) - demonstra como cada algoritmo captura aspectos diferentes dos mesmos dados. O KMeans revela a estrutura temporal-duração, enquanto o DBSCAN adiciona a dimensão da classificação indicativa, oferecendo uma perspectiva tridimensional mais rica.
Esta complementaridade dos algoritmos valida a abordagem metodológica adotada, onde diferentes técnicas de clustering revelam facetas distintas da estratégia de curadoria da Netflix.

4.5 Implicações Estratégicas

Os resultados sugerem que a Netflix adota uma estratégia de "núcleo e periferia", onde um núcleo substancial de conteúdo mainstream atende à maioria da audiência, complementado por conteúdo diversificado que atende nichos específicos e contribui para a percepção de um catálogo abrangente.
Os outliers identificados pelo DBSCAN, embora representem uma minoria do catálogo, podem ter valor estratégico desproporcional ao seu número. Eles servem para diferenciar a plataforma, atender públicos específicos, e manter a percepção de diversidade e qualidade do catálogo.

5. Limitações do Estudo

É importante reconhecer as limitações desta análise. Primeiro, o dataset utilizado contém apenas uma amostra do catálogo Netflix até abril de 2021, não refletindo mudanças posteriores ou variações regionais no conteúdo disponível. Segundo, foram utilizados apenas três variáveis na análise, enquanto decisões estratégicas de curadoria certamente envolvem múltiplos fatores não capturados nos dados.
Terceiro, a análise não considera dados de engajamento dos usuários, audiência, ou performance comercial dos títulos, fatores que são fundamentais para compreender o verdadeiro valor estratégico de diferentes segmentos do catálogo. Quarto, a transformação da classificação indicativa em escala numérica, embora necessária para a análise, pode não capturar adequadamente as nuances regulatórias e culturais dessas classificações.

6. Conclusão

A aplicação de técnicas de clusterização sobre o catálogo de filmes da Netflix demonstrou ser uma abordagem eficaz para identificar padrões estratégicos na curadoria de conteúdo. O KMeans revelou uma estrutura de portfólio balanceada com quatro segmentos distintos, enquanto o DBSCAN evidenciou a concentração de densidade no conteúdo mainstream e destacou a importância do conteúdo atípico.
Ambos os métodos confirmaram uma tendência clara: o foco principal da Netflix está em filmes lançados na última década, com duração média e classificação intermediária, atendendo às preferências da audiência contemporânea. No entanto, a manutenção de conteúdo clássico e experimental demonstra uma estratégia sofisticada de diferenciação e atendimento a múltiplos segmentos de mercado.
Os outliers identificados pelo DBSCAN, embora numericamente minoritários, representam uma parcela estrategicamente importante do catálogo, contribuindo para a diversidade e unicidade da oferta da plataforma. Esta descoberta sugere que a análise de conteúdo atípico pode ser tão importante quanto a compreensão dos padrões dominantes.

Referências

[1] TidyTuesday. Netflix Titles Dataset. Disponível em: https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-04-20 . Acesso em: jul. 2025.
[2] ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. 1996.
[3] LLOYD, S. Least squares quantization in PCM. IEEE Transactions on Information Theory, 1982.
[4] PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2011.
[5] JAIN, A. K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 2010.