Descoberta de Subgrupos Aplicada ao Mercado de Jogos Digitais: Exploração de métricas delimitadoras de subgrupos em jogos positivamente avaliados na plataforma Steam

Gabriel B. Freddi[1], João Vitor S. Depollo[2], Pedro O. Guedes[3] Tarcízio A. S. Lafaiete[4], Vinícius A. F. Resende[5]

[1]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

gabrieloriginal@ufmg.br, pedro-og2002@ufmg.br, jvsd@ufmg.br, tarcizio-augusto@ufmg.br, viniciusfariaresende@ufmg.br

Abstract. *This study investigates a publicly available dataset from the Steam platform, accessed through Kaggle and compiled using the Steam API and Steam Spy. The proposed approach employs techniques such as SDMap and MCTS to identify subgroups with more positive reviews and, with the team's expertise, analyze the discovered rules to gain relevant insights in the context of the gaming industry.*

Resumo. *Este estudo investiga um conjunto de dados disponível publicamente da plataforma Steam, acessado através do Kaggle e compilado utilizando a API da Steam e o Steam Spy. A abordagem proposta utiliza técnicas como SDMap e MCTS para identificar subgrupos com mais avaliações positivas e, com o conhecimento da equipe, analisar as regras encontradas para obter informações relevantes no contexto da indústria de jogos.*

1. Introdução

O trabalho em questão explora uma base de dados contendo informações diversas sobre o vasto catálogo de jogos da Steam, a maior plataforma de jogos para computador. A base de dados está disponível publicamente na plataforma Kaggle[6], e consta que a coleta de dados foi feita pela própria API da Steam, bem como pela utilização do serviço Steam Spy.

O mercado de jogos digitais é extremamente relevante no contexto global, sendo que em 2022 a indústria de jogos movimentou um total de USD 217.06 bilhões. Além disso, existe uma expectativa de crescimento anual de 13.4% entre os anos de 2023 e 2030, fazendo com que o tamanho de mercado dos videogames seja estimado em USD 583.69 bilhões no ano de 2030[5]. Mais especificamente sobre a Steam, faz-se necessário considerar a relevância dos videogames no contexto de jogos de PC. Segundo um levantamento do site Statista, estima-se um total de 1.86 bilhões de jogadores de PC no mundo no ano de 2024, sendo que foi observado um aumento de 69.09% entre os anos de 2008 e 2024[9].

Considerando a relevância dos jogos digitais de computador na contemporaneidade, evidencia-se o valor de uma exploração sistemática e sofisticada dos dados descritivos relacionados com o objetivo de verificar certos comportamentos dos usuários em relação aos jogos disponíveis. Utilizando técnicas modernas de aprendizado descritivo, pode-se colher informações relevantes no que tange à segmentação de características comuns entre jogos bem avaliados no marketplace da

Steam. Dessa forma, essas informações podem ser utilizadas a fim de guiar empresas de desenvolvimento de jogos na confecção de videogames com maior possibilidade de adesão pelos usuários da plataforma Steam.

Com o objetivo supracitado em mente, é plausível conceber a utilização de descoberta de subgrupos como forma de atacar o problema. Com a utilização destas ferramentas, é possível realizar a identificação de subgrupos similares. A similaridade em questão, entre os subgrupos, é avaliada quanto a algum atributo alvo retirado da base de dados. Considerando a característica dos métodos supervisionados, de escolha de atributos de interesse, podemos contemplar características de relevância comercial. Por isso, observar subgrupos formados por jogos bem avaliados segundo os usuários, pode evidenciar atributos comuns entre jogos que são sucesso de vendas.

A posse de informações descritivas acerca do catálogo de jogos da Steam é de grande valia tanto para as produtoras de jogos que já comercializam seus produtos na plataforma quanto para empresas que ainda não usam o serviço. Os dados indicam um perfil de preferência dos usuários da Steam, revelando características desejáveis dos produtos a serem oferecidos, aumentando assim a possibilidade de um lançamento lucrativo na plataforma.

Em suma, o artigo aqui presente propõe a utilização de descoberta de subgrupos como ferramenta de planejamento estratégico no fluxo de desenvolvimento e venda de jogos na Steam. Como metodologia para esse processo serão utilizados os algoritmos SDMap e MCTS, tendo os resultados obtidos apresentados bem como a análise destes.

2. Materiais e Métodos

A descoberta de subgrupos (Subgroup Discovery) é uma área da mineração de dados que busca identificar subgrupos específicos dentro de um conjunto de dados que possuem características ou comportamentos notáveis em relação a uma medida de interesse. Dois métodos importantes nessa área são o SDMap com Beam Search[3] e o Monte Carlo Tree Search (MCTS)[2]. O SDMap utiliza a Beam Search para explorar eficientemente o espaço de busca, expandindo apenas um subconjunto dos subgrupos mais promissores em cada iteração, o que otimiza a descoberta de subgrupos relevantes sem explorar todas as combinações possíveis. Por outro lado, o MCTS combina busca aleatória e sistemática, balanceando a exploração de novos subgrupos com a exploração de subgrupos conhecidos, sendo particularmente útil em espaços de busca grandes e complexos.

Para avaliar a qualidade dos subgrupos encontrados, várias métricas são utilizadas, sendo a Weighted Relative Accuracy (WRAcc)[7] e a métrica Beta[8] duas das mais importantes. A WRAcc mede o ganho relativo em relação à distribuição esperada, considerando a proporção de instâncias no subgrupo e sua precisão em comparação com a média do conjunto de dados completo, penalizando subgrupos muito pequenos ou muito grandes. Isso garante que os subgrupos identificados sejam tanto significativos quanto representativos. A métrica Beta, por sua vez, ajusta a importância da precisão e da cobertura através de um parâmetro, permitindo balancear a descoberta de subgrupos de acordo com diferentes necessidades analíticas.

Esses métodos e métricas são cruciais para a descoberta eficaz de subgrupos relevantes em grandes conjuntos de dados, proporcionando insights valiosos em diversas aplicações. Ao utilizar o SDMap com Beam Search ou o MCTS, e ao avaliar os subgrupos com métricas como WRAcc e Beta, é possível identificar subgrupos que não só se destacam estatisticamente, mas também são significativos para a compreensão e exploração dos dados.

2.1. A Base de Dados

Como dito no tópico anterior, a base de dados utilizada na avaliação foi retirada do site Kaggle[6]. Estes dados dispostos no site, foram coletados utilizando um *web Scraper* desenvolvido pelo *game developer* Martin Bustos que se encontra disponível no github para uso público. Este *web scraper* faz uso de apis da steam para visitar as páginas de jogos presentes na plataforma e coletar os dados salvando-os em dois formatos: um .csv e um .json.

Estes dados foram coletados no período entre 23/04/2022 até 04/07/2023, durante esta pesquisa foram coletados dados de mais de 85 mil jogos. Para cada um destes jogos presentes no conjunto existem 39 aspectos diferentes registrados para cada, alguns destes atributos são: Nome, Data de lançamento, Número estimado de donos, Gêneros, Tags, Categorias e Preço.

Devido a natureza da coleta dos dados que foi realizada utilizando um *web scraper* é comum que haja problemas na conjectura de dados com a existência de valores nulos ou mal-formatados. Além disso, nem sempre os dados estão em um formato ideal para realizar a descoberta de subgrupos, por isso se torna necessário a limpeza, tratamento e processamento dos dados antes da utilização de algoritmos em cima desta base.

2.2. Processamento dos Dados

A primeira limpeza de dados realizada foi relacionada com os nomes de jogos, uma vez que, haviam na base de dados, itens sem nome e com valores duplicados. Os jogos sem nome foram removidos devido a dificuldade de realizar melhores observações dentro dos subgrupos descobertos, já as duplicatas foram retiradas preservando a cópia com o maior valor da coluna de *Estimated Owners*.

Em seguida, foi realizada uma segunda limpeza de dados, na qual diversas colunas foram deletadas no conjunto de dados. Este procedimento foi dividido em três etapas, sendo que na primeira foram removidos os atributos com um elevado número de valores nulos, sendo estes: *Score Rank, Metacritic url, Notes, About the Game, Screenshots, Movie* e *Developers.*

Adiante as colunas que foram determinadas como não interessantes, de difícil proveito ou redundantes em relação a outros dados foram excluídas. As propriedades removidas foram: *Supported languages, Full audio languages, Recommendations, Windows, Mac, Linux, Tags* e *Header Images.*

Por fim na terceira etapa de limpeza as colunas: *Average playtime 2 weeks, Median playtime 2 weeks, Average playtime forever, Median playtime forever, Achievements, Required age* e *Price,* foram retiradas do conjunto devido a sua má formatação e devido à inconsistências nos dados percebidas através de análises manuais em relação aos dados disponíveis na plataforma.

Nas demais colunas algumas transformações foram necessárias adequações dos dados para seu uso nos algoritmos de descoberta de subgrupos. As colunas *Review, Support email* e *Support url* tiveram suas informações convertidas em formato booleano indicando a existência de dados ou não nas linhas.

Os atributos *Genres* e *Categories* foram convertidas em um formato one hot encode, contudo após uma análise de frequência dos elementos presentes nestes, foram removidos os rótulos presentes nos extremos da distribuição observada. Em *Release Date*, a sua representação foi convertida para uma coluna representando o ano de lançamento e quatros colunas booleanas representando a temporada em que o jogo foi lançado.

Finalizando o tratamento de dados, houve uma binarização da coluna *Publishers*. Neste procedimento, com a utilização de dados do *SteamDB*[10], as publicadoras foram divididas entre famosas ou não a partir da sua colocação na lista. Com isto, tem-se resultados booleanos para o atributo.

2.3. SD–Map

Considerando o algoritmo SD-Map descrito em Atzmueller, M., Puppe, F. (2006)[1], e a implementação fornecida pela biblioteca Python *pysubgroup*[4]. Fomos capazes de executar o algoritmo SD-Map, utilizando a estratégia de busca de Beam Search[3], expandindo apenas um subconjunto dos subgrupos mais promissores em cada iteração. Após a eficiente limpeza dos dados, especialmente de colunas com um espaço amostral muito grande de valores (e.g *Name, Release Date*), foi possível executar eficientemente o algoritmo com os dados.

Podendo executar o algoritmo livremente, variamos constantemente o atributo alvo, de forma a cobrir uma grande quantidade de abordagens diferentes que poderiam revelar subgrupos diversos e não triviais. No que tange aos atributos alvos, começamos com uma classificação simples que selecionava os jogos que tinham uma avaliação muito positiva seguindo a nota do Metacritic[1]. Posteriormente, seguiu-se para uma abordagem que considerava primariamente as avaliações na própria plataforma Steam, inicialmente considerando jogos que apresentaram um superávit de 10% nas avaliações positivas em relação às negativas, e, por fim, subimos o valor para 20%.

Concomitantemente com as variações nos atributos alvos, foram variados também os parâmetros do algoritmo, especialmente a quantidade de subgrupos buscados e, principalmente, a profundidade de busca. Este segundo parâmetro representou mudanças significativas nos resultados obtidos e na qualidade dos subgrupos, mesmo que chegando a um platô em determinado momento. Variando o atributo alvo e os

[1] Metacritic é um site que compila críticas de filmes, jogos, músicas e etc. e gera uma nota a partir da média ponderada das notas agregadas.

parâmetros do algoritmo, vários resultados diferentes foram observados ao decorrer dos testes.

De modo geral, os testes mostraram subgrupos coerentes, no sentido de que facilmente entendia-se a relação entre os atributos obtidos que classificavam o subgrupo. Ademais, nas iterações finais, notou-se que a métrica de qualidade utilizada apresentava bons resultados, para a avaliação da qualidade dos grupos utilizou-se o WRAcc[7], que também é mensurado pela implementação da própria biblioteca do *pysubgroup*.

Contudo, após a avaliação dos subgrupos notou-se que, apesar da coerência entre os grupos, as informações enaltecidas não eram significativas e não foram consideradas relevantes para o propósito da pesquisa de forma geral. Isso porque, os subgrupos no geral eram desinteressantes ou óbvios, no sentido de que não acrescentavam novidade no assunto tratado, indo de forma contrária ao objetivo da pesquisa.

Por fim, considerando os resultados obtidos pelo SD-Map em contraste com os obtidos pelo uso do MCTS, definiu-se que os resultados do primeiro não tinham um valor intrínseco comparável. Isso ruminou na opção de descartá-lo como ferramenta para a obtenção de resultados.

2.4. MCTS

Utilizando a implementação Guillaume(2017)[2], feita em Java , é possível rodar o MCTS com diversos parâmetros e ajustá-los para melhor servir ao objetivo. Como configuração, é possível escolher entre várias medidas de qualidade dos subgrupos, como WRAcc , F1, RelativeF1, WeightedRelativeF1, WKL, FBeta, RelativeFBeta, WeightedRelativeFBeta,RAcc, Acc, HammingLoss, ZeroOneLoss, ContingencyTable, Jaccard e Entropy. Como *target*, o grupo utilizou:

$$Target = Positives > Negatives * 1.2 \wedge Positives > AVG_{Positives}$$

Equação 1: Target.Quantidade de avaliações positivas são no mínimo 120% das avaliações negativas e quantidade de positivas é maior que a média total de notas positivas.

Por padrão, o algoritmo utiliza o *WRAcc* como métrica, mas ao rodar o algoritmo por diversas vezes e ajustando parâmetros como redundância, número de interações, suporte mínimo e entre outros, o *WRAcc* retornava subgrupos muito imprecisos e redundantes, como: $1998 \leqslant ano \leqslant 2023$.

Além disso, a documentação não deixa muito claro quais parâmetros poderiam ser ajustados a fim de reduzir essa questão. Dessa forma, com o objetivo de aumentar a precisão dos subgrupos, ao avaliar entre diversas métricas possíveis, optou-se por utilizar o *FBeta*, que permite ponderar sobre a precisão e o *recall*. Nesse sentido, a implementação disponibiliza 2 parâmetros em relação ao *FBeta*, o *xBeta* e o *lBeta*,

ambos não possuem a documentação de como irão impactar diretamente na qualidade. Estudando o código, entende-se que o *xBeta* e o *lBeta* são subdivisões do *FBeta*, em que o *xBeta* pesa a precisão e o *lBeta* o recall:

$$\beta = 0.5 * (1 + tanh(Global.xBeta - suppL)/lBeta)$$

Equação 2: Cálculo do Beta no MCTS utilizando *FBeta* como medida de qualidade

Além disso, o programa só aceita um tipo de tipo de atributo, no *dataset* estudado foi utilizado o atributo numérico. Dessa forma, também foi necessário mapear booleanos para 0 ou 1.

A execução é relativamente simples de ser executada, após o tratamento, basta informar a pasta em que os inputs se encontram e a pasta de output. Para esta implementação, são necessários 3 arquivos de inputs. Os dados no formato arff, que descreve uma lista de elementos que compartilham um grupo de atributos e descrevem o *dataset*, um csv de qualidades, com uma coluna falando quais elementos atingiam a qualidade necessária e outro arquivo com todo o *dataset*.

2.4.1 Parâmetros

Tipo de Atributos	
Número de Interações	10000
Min Supp	10
Max Redundancy	0.6
Measure	FBeta
Xbeta	5
LBeta	70

Tabela 1: Parâmetros utilizados para rodar o MCTS4DM

3. Resultados

Assim como foi citado anteriormente, os conjuntos de descritores obtidos ao longo deste trabalho foram minerados com o algoritmo MCTS4DM. Nas subseções seguintes, cada um dos conjuntos encontrados será abordado separadamente fazendo

considerações sobre a regra obtida e o que ela representa em relação ao universo de jogos. Os títulos das subseções são idealizados pelos autores, buscando representar em poucas palavras o que os descritores estão delimitando.

3.1 Jogos de sucesso contra a crítica

Os descritores obtidos podem ser vistos na Tabela 2. É possível notar uma certa redundância nas regras obtidas pelo algoritmo, já que ele cita três intervalos separados de quantidade estimada de jogadores, sendo que essa é uma regra auto excludente, ou seja, se um jogo está em um intervalo, ele não estará em mais nenhum outro. O intervalo de pontuação Metacritic é relativamente baixo, principalmente quando comparado a jogos AAA[2], que costumam receber pontuação acima de 90. Apesar disso, todos os jogos deste subgrupo chegam a pelo menos 2.000.000 de jogadores, o que menos de 0,04% dos jogos da base conseguiram alcançar.

51.0 <= Metacritic score <= 74.0	Categories_Steam Leaderboards = 1.0	Estimated owners_100000000 - 200000000 = 0.0
Estimated owners_2000000 - 5000000 = 1.0	Estimated owners_0 - 20000 = 0.0	

Tabela 2: Regras delimitadoras do primeiro subgrupo.

Outro descritor que pertence ao conjunto é o que indica a presença da funcionalidade "*Steam Leaderboards*[3]" em todos os grupos, que é uma tabela que mostra a pontuação dos jogadores em relação aos amigos e outros jogadores mundiais. Esse é um recurso que trabalha com a ideia de comunidade e competitividade nos jogos, ainda que eles sejam de jogador único, como motivador para o público se manter engajado. O "Steam Leaderboards" é uma funcionalidade que pode se mostrar relevante para desenvolvedores na busca por jogos que terão mais popularidade e engajamento dos usuários ativos, permitindo troca entre os jogadores e potencialmente aumentando a quantidade de compras.

[2] Jogos triplo A (AAA) são os com os maiores níveis de orçamento e divulgação em relação aos pares no mercado.

[3] Os *Steam Leaderboards* são tabelas persistentes com entradas ordenadas automaticamente e podem ser usadas para exibir tabelas globais e de amigos no seu jogo e na página da sua comunidade.

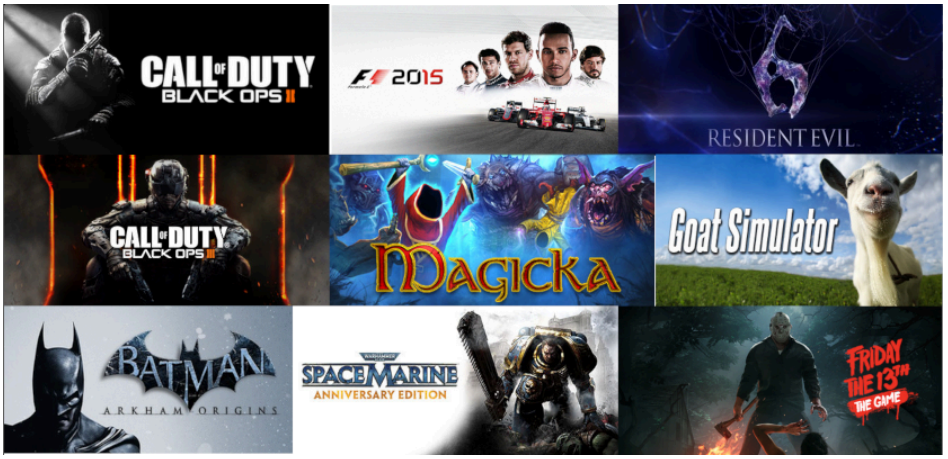

Figura 1: Exemplos de jogos do subgrupo **3.1**.

3.2 Jogos de dedicação exclusiva

Os jogos delimitados pelo segundo conjunto de descritores foram todos publicados por empresas consideradas conhecidas, com a funcionalidade de múltiplos jogadores e o *Steam Workshop*, que facilita a criação e compartilhamento de modificações do jogo entre os jogadores. A partir dessas informações, é possível deduzir que foram jogos com bom suporte de divulgação pelas empresas publicadoras, além de terem funcionalidades que levam os jogadores para um maior senso de comunidade, podendo jogar com amigos e personalizar o conteúdo base facilmente. Características como essas levam os analistas a acreditar que o público detentor dos títulos seria bastante extenso, o que não é verificado em análises posteriores.

470.0 <= Peak CCU <= 2650.0	is_publisher_known = 1.0	Categories_Multi-player = 1.0
Categories_Steam Workshop = 1.0	Estimated owners_2000000 - 5000000 = 0.0	

Tabela 3: Regras delimitadoras do segundo subgrupo.

Apesar da maioria dos títulos terem entre 1.000.000 e 2.000.000 de cópias obtidas por jogadores, a quantidade de jogadores simultâneos ainda é baixa quando comparada a demais jogos também publicados por empresas conhecidas. Enquanto existem mais de 130 jogos de publicadoras conhecidas que obtiveram mais de 2.650 jogadores simultâneos, quando os descritores obtidos são aplicados, são encontrados apenas 16 com os descritores do subgrupo que superaram essa marca.

Esse comportamento pode ser explicado pelo fato de que neste subgrupo são encontrados jogos de estratégia e simulação, que demandam muito tempo e dedicação do jogador que precisa pesquisar sobre características específicas do produto, como

segredos, técnicas, itens, regras, otimizações, personalizações, mods, entre outros. Outro fator importante a se observar em relação a este gênero de jogos é o peso que as publicadoras dão para a credibilidade, visto que devido a complexidade e tempo gastos para se aprimorar nestes jogos, a tendência dos jogadores é priorizar jogos de publicadoras mais conhecidas e que possuem comunidades pré estabelecidas em volta destas.

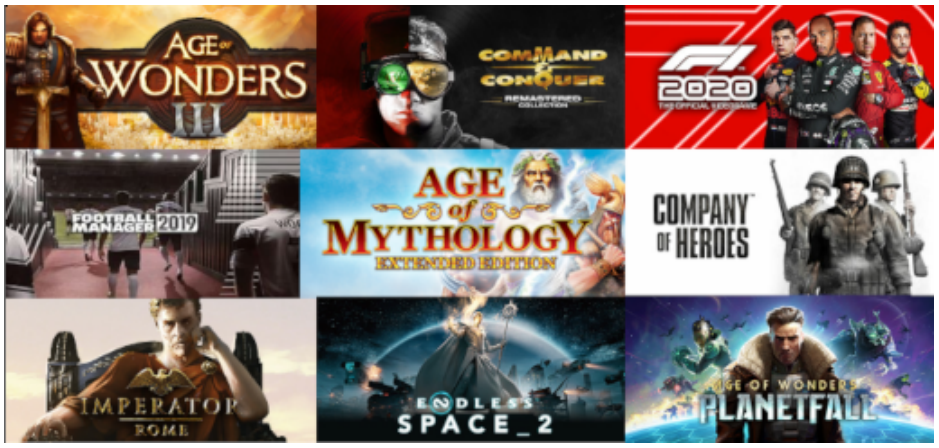

Figura 2: Exemplos de jogos do subgrupo **3.2**.

4. Conclusões

Descoberta de subgrupos (SD)[1] é um campo ativo de pesquisa atualmente , com diversos avanços e novas técnicas sendo desenvolvidas para diversas aplicações. Este artigo visou utilizar algoritmos conhecidos de SD em cima da base de jogos da Steam para encontrar regras e conjuntos interessantes de jogos que possuem uma relativa aceitação do público e entender o que estes grupos sinalizam em relação ao ecossistema da Steam e suas publicadoras de jogos.

Em linhas gerais, a descoberta de subgrupos para este *dataset* necessitou de vários tratamentos e configurações para a utilização dos dados como fonte do modelo. Além disso, os dois algoritmos apresentados possuem em sua implementação limitações no que tange a formatação da entrada, com destaque para o mcts4dm que obriga a conversão dos dados para um tipo único na entrada, sendo assim adicionando mais uma limitante ao uso dos dados da base selecionada.

Apesar destas pontuações anteriores, os resultados obtidos se apresentaram motivantes e abre espaço para investigações futuras sobre a temática, visto que ainda há limitações nos dados coletados, tanto na quantidade de dados utilizados, quanto na qualidade destes, visto que muitas colunas apresentaram problemas com dados nulos, mal-formatados e incongruentes.

Desta forma, o horizonte se encontra aberto para novos projetos com perspectivas interessantes, observando não apenas os dados da Steam mas também de plataformas que auxiliam na divulgação destes jogos como por exemplo o Youtube e a Twitch para entender a influência destes na percepção do público em relação a aceitação do público.

Refências

1. Atzmueller, M., Puppe, F. (2006). SD-Map – A Fast Algorithm for Exhaustive Subgroup Discovery. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds) Knowledge Discovery in Databases: PKDD 2006. PKDD 2006. Lecture Notes in Computer Science(), vol 4213. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11871637_6.
2. Bosc, G., Boulicaut, JF., Raïssi, C. et al. Anytime discovery of a diverse set of patterns with Monte Carlo tree search. Data Min Knowl Disc 32, 604–650 (2018). https://doi.org/10.1007/s10618-017-0547-5.
3. Dragan Gamberger and Nada Lavrac. 2002. Expert-guided subgroup discovery: methodology and application. J. Artif. Int. Res. 17, 1 (July 2002), 501–527. https://dl.acm.org/doi/10.5555/1622810.1622825.
4. GITHUB. **pysubgroup**. Disponível em: https://github.com/flemmerich/pysubgroup. Acesso em: 30 jul. 2024.
5. GRAND VIEW RESEARCH. **Video Game Market Size, Share & Trends Analysis Report By Device (Console, Mobile, Computer), By Type (Online, Offline), By Region (Asia Pacific, North America, Europe), And Segment Forecasts, 2023 - 2030**. Disponível em: https://www.grandviewresearch.com/industry-analysis/video-game-market. Acesso em: 30 jul. 2024.
6. KAGGLE. **Steam Games Dataset**. Disponível em: https://www.kaggle.com/datasets/fronkongames/steam-games-dataset/data?select=games.csv. Acesso em: 30 jul. 2024.
7. Lavrac, N., Flach, P., Zupan, B.: Rule evaluation measures: a unifying view. In: Dzeroski, S., Flach, P. (eds.) ILP 1999. LNCS (LNAI), vol. 1634, pp. 174–185. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-48751-4_17.
8. R. Baeza-Yates and B. Ribeiro-Neto (2011). Modern Information Retrieval. Addison Wesley, pp. 327-328.
9. STATISTA. **Number of PC gaming users worldwide from 2008 to 2024**. https://www.statista.com/statistics/420621/number-of-pc-gamers/. Acesso em: 30 jul. 2024.
10. STEAMDB. **Popular Game Publishers on Steam**. Disponível em: https://steamdb.info/publishers/. Acesso em: 30 jul. 2024.