**Trabalho de Conclusão de Curso**

PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

**ALUNO: Murilo Wolfart**

**ORIENTADOR: Flávio Arthur Leal Ferreira**



Sumario

´

[**1. RESUMO 2**](#_Toc162986287)

[**2. INTRODUÇÃO 3**](#_Toc162986288)

[**3. TRABALHOS RELACIONAIS 4**](#_Toc162986289)

[**4. METODOLOGIA 6**](#_Toc162986290)

[**5. RESULTADOS 9**](#_Toc162986291)

[**6. DISCUSSÃO 13**](#_Toc162986292)

[**7. CONCLUSÃO E TRABALHOS FUTUROS 14**](#_Toc162986293)

[**8. REFERÊNCIAS 15**](#_Toc162986294)

**ARTIGO CIENTÍFICO**

# RESUMO

A fim de descobrir formas de atrair mais usuários para jogos na App Store, usou-se um dataset do Kaggle contendo 2000 jogos junto às suas descrições e avaliações para fazer uma análise e treinamento de um modelo e verificar se existe alguma relação entre o conteúdo do jogo, descrito no campo “descrição”, e as avaliações do mesmo. Através deste dataset descobriu-se que existe uma quantiade enorme de avaliações positivas (4 estrelas ou mais), demonstrando alta satisfação dentre os usuários que instalam e avaliam o jogo, e portanto não produzindo um modelo eficaz o suficiente para prever avaliações positivas. No entanto, aproximadamente metade do dataset contém menos de 30.000 avaliações feitas, e um quarto contém menos de 5.300 avaliações. Dessa forma, treinou-se um modelo a fim de verificar relações entre o conteúdo (descrição) de um jogo e a quantidade de avaliações feitas (gerando portanto um maior engajamento). Analisando as tabelas TF-IDF geradas para duas classes distintas “highRatingCount” e “lowRatingCount” notou-se, dentre os dados de maior quantidade de avaliações, maior presença de palavras como “subscription” (assinatura), “slot” (caça-níqueis), “free” (gratuito), “trial”, “casino”, “battle” (batalha), além de uso de estrutura em tópicos, indicando um maior engajamento para estes temas e para descrições mais claras, objetivas e detalhadas. O modelo treinado previu corretamente as classes numa taxa de aproximadamente 70% (acurácia), indicando portanto uma relação consideravelmente forte entre descrição de jogos e quantidade de avaliações.

# INTRODUÇÃO

No atual universo de dispositivos móveis, a App Store, loja digital da Apple, desempenha um papel fundamental na disponibilização e lançamento de aplicativos. Atualmente existem mais de 1.8 milhões de apps disponíveis na plataforma dividos em diversas categorias, cada uma com seu perfil de usuário e particularidades. Dentre estas categorias, está a de jogos digitais, que vem crescendo cada vez mais ao longo dos anos. Para os desenvolvedores, um dos desafios no lançamento de um jogo é obter avaliações positivas e atrair usuários, uma vez que as únicas informações que estes terão inicialmente é a descrição, categoria e algumas capturas de tela antes de instalá-lo. Nesse contexto, torna-se útil estudar formas de atrair jogadores utilizando estes pequenos espaços na página de cada aplicativo.

Este trabalho propõe-se a analisar a relação entre descrições de jogos e avaliações de usuários na App Store, se ela existe e como influencia na popularidade do jogo. Por meio de análises estatísticas e modelagem preditiva, procura-se identificar padrões e correlações que possam oferecer informações valiosas para desenvolvedores na criação e promoção de jogos, bem como temas, características e funcionalidades que atraem mais atenção.

O modelo preditivo construído neste estudo também é utilizado para, dado uma descrição de um jogo, predizer se ele irá obter uma boa quantidade de avaliações na loja, atraindo mais usuários e gerando um maior engajamento. A aplicação deste modelo pode auxiliar desenvolvedores na otimização de suas estratégias de marketing e na criação de conteúdo mais atrativo para usuários da App Store.

Ao compreendermos melhor como as descrições dos jogos se relacionam com o engajamento dos usuários e quais características influenciam positivamente na popularidade de um jogo, podemos não apenas melhorar a visibilidade e o sucesso dos jogos na plataforma, mas também contribuir para uma melhor experiência para os jogadores.

# TRABALHOS RELACIONAIS

Um trabalho semelhante foi feito por Aisulu Omar e publicado no website Medium.com, onde vários aplicativos na Play Store (de Google) são analisados, a fim de verificar quais das informações básicas como ano de lançamento, categoria e preço influenciam na avaliação de usuários. O trabalho utiliza um dataset disponível no Kaggle contendo informações de 10 mil aplicativos e diferentes regressores, fundamentalmente utilizando o regressor XGBoost para geração do modelo final que prevê avaliações de um aplicativo com base nas informações básicas. Este regressor também é utilizado para verificar que aplicativos das categorias esportes, utilitários e também aplicativos gratuitos obtem uma melhor avaliação num geral.

Em 2019 um artigo feito por Kevin Daimi e Noha Hazzazi pela universidade de Detroit analisou diversos aplicativos na App Store a fim de buscar relações entre avaliações de usuarios e informações básicas, muito semelhante ao trabalho feito por Omar. Neste trabalho, somente é utilizado para predizer a avaliação média dos aplicativos os atributos quantidade de avaliações, número de dispositivos suportados, número de idiomas suportados e número de capturas de telas fornecidas. O trabalho visa utilizar diferentes técnicas de aprendizado de máquina para predizer as avaliações médias dos aplicativos, como SVM, florestas aleatórias e regressão linear, sendo árvores aleatórias o algoritmo que gerou os melhores resultados.

Um artigo de 2018 publicado por Maleknaz Nayebi busca correlacionar avaliações de 30793 aplicativos na App Store com relatos feitos no Twitter, a fim de descobrir padrões e verificar o quão útil pode ser o feedback coletado na rede social, quando comparado às avaliações feitas na loja. Resultados desta pesquisa mostram que os tweets de modo geral permitem identificar de forma eficiente não só o sentimento de um usuário para com um aplicativo, mas também outras informações valiosas como pedidos de funcionalidade e relatos de bugs.

Em 2013 um estudo foi realizado por Bin Fu *et al*. em que uma análise a três níveis de detalhamento diferentes é feita em cima de aplicativos na Google Play Store. Através desta pesquisa foi possível identificar motivos pelos quais os usuários não gostam de um aplicativo, além de inconsistências nas avaliações de usuários e preferências gerais por tipos de apps específicos e trends em alta.

Em 2014 um artigo publicado por Mir Riyanul Islam observa grandes diferenças entre as avaliações numéricas de usuários e as avaliações escritas, para diversas aplicações disponíveis na Google Play Store. A fim de remover esta ambiguidade, o estudo propõe criar uma nova métrica calculada a partir das duas informações, que possa ser utilizada como avaliação (escore) final de um aplicativo. Neste trabalho, uma análise de sentimento também é utilizada nos corpus de escrita das avaliações disponíveis a fim de calcular a nova métrica.

Por fim, um estudo semelhante ao descrito aqui foi feito por Dayi Lin *et al.* em 2018, utilizando avaliações de jogos na plataforma Steam a fim de detectar correlações com avaliações de jogos de dispositivos móveis e, portanto, entender se é possível utilizar das mesmas descobertas em ambas. As avaliações feitas na plataforma também são analisadas para verificar se é possível extrair tópicos relevantes e percepções dos usuários num geral. O estudo conclui que as avaliações feitas para jogos na plataforma são bem diferentes de avaliações de jogos para dispositivos móveis, citando a quantidade de horas jogadas como um parâmetro importante que é inexistente nestes últimos.

# METODOLOGIA

Para realizar o estudo, foi utilizado um dataset contendo informações de 2000 jogos na App Store, disponível no Kaggle. O dataset contém diversas informações sobre os apps, incluindo descrição, autores, avaliação média, número de avaliações, gênero primário, preço, notas de lançamento, dentre outros. Além disso, utilizou-se de Python e suas bibliotecas para as análises, transformações e treinamento de modelo descritos neste artigo.

Através das informações do dataset, convertidas para um dataframe da biblioteca *pandas*, o primeiro passo foi identificar por ausência de dados. Foram constatadas 58 lacunas no campo “preço” e 31 no campo “notas de lançamento”, com os outros campos nunca estando vazios. No entanto, foi também constatado que dos jogos que possuem preço, apenas um efetivamente possui um custo de instalação, enquanto 1941 são todos gratuitos para um primeiro download. Além do preço, também foi feita uma análise sobre os dados de avaliação média dos usuários, número de avaliações feitas e descrição. Abaixo é exibida duas imagens referentes à distribuição de dois campos do dataset: avaliação média (à esquerda) e número de avaliações (à direita, com extração de alguns outliers para melhor visualização).

|  |  |
| --- | --- |
| A graph of a number of blue lines  Description automatically generated | A graph of a number  Description automatically generated |

Com base nessa análise exploratória é possível visualizar que, dos 2000 jogos, 1879 possuem avaliação média superior a 4 estrelas num total de 5, e que 1451 possuem avaliação superior a 4.5 estrelas. Mais precisamente, a mediana do campo avaliação média fica em 4.62, e o primeiro quartil fica em 4.47. Tais informações demonstram uma satisfação geral dentre usuários que instalam e avaliam os jogos. Para o número de avaliações, a distribuição é um pouco mais uniforme, variando entre 0 e 10 milhões, com mediana em 26 mil, primeiro quartil em 5 mil e terceiro quartil em 108 mil, embora ela ainda siga uma distribuição semelhante à exponencial negativa e possua alguns outliers. Quanto ao campo descrição, a média da quantidade de palavras fica em 300, com mediana em 279 e primeiro e terceiro quartis em 159 e 428 respectivamente.

Após a análise inicial, aplicou-se um algoritmo de stemming e remoção de stop words com auxílio da biblioteca *ntlk*, utilizando *Porter Stemmer* para a realização do primeiro. Feito isso, gerou-se uma matriz de palavras e suas quantidades por linha através do algoritmo *CountVectorizer*, pertencente à biblioteca *SciKit*. Além disso, também foi utilizado um *TfidfTransformer* para realçar palavras que possuem mais importância no contexto de descrição do aplicativo.

Com as transformações necessárias feitas, duas sequências de testes foram feitas à partir de duas novas propriedades: “goodRating”, que determina se um jogo possui uma avaliação boa, e “highEngagement”, que determina se um jogo irá possuir uma quantidade significativa de avaliações de usuário. Para o primeiro caso, utilizou-se inicialmente como critério de avaliação boa o limiar de 4.5 estrelas, tendo em vista a alta distribuição de valores acima de 4 para esta informação, e eventualmente testando valores novos. Já para o segundo caso, utilizou-se 30.000 avaliações recebidas como critério, um valor próximo da mediana desta coluna. Estas duas propriedades seriam então as classes preditas pelo nosso modelo treinado.

Para treinamento de modelo utilizou-se dois algoritmos: *Support Vector Machines* e também *Histogram Gradient Boosting*, ambas de *SciKit*. 20 modelos foram gerados para diferentes combinações de conjuntos de treinamento e teste, onde o conjunto de testes sempre era 20% do conjunto original. Ao final computou-se, através da matriz de confusão, a média dos valores de acurácia, precisão e recall.

Além do modelo criado, também foi-se feita uma análise das palavras de maior incidência dentre jogos com maior engajamento, verificando os maiores valores numéricos nas tabelas TF-IDF. Mais precisamente, separou-se os dois conjuntos conforme a classe e, após computada a matriz TF-IDF em ambos, subtraiu-se a segunda (jogos com baixo engajamento) da primeira (jogos com alto engajamento), para eliminar artigos e palavras que não possuem tanta importância, listando assim as palavras que tendem a garantir mais número de avaliações dos usuários.

# RESULTADOS

Para a propriedade “goodRating”, representando se um jogo obteve uma avaliação positiva ou não, os modelos preditos utilizando *Support Vector Machines* atingiram uma média de valores de acurácia, precisão e recall de 77.31%, 77.48% e 97.05% respectivamente. Mais precisamente, os modelos previram em média 282 verdadeiros positivos, 82 falsos positivos, 9 falsos negativos e 27 verdadeiros negativos. Nestes casos o critério utilizado fora de 4.5 estrelas de 5. Embora os valores relativos pareçam bons, a quantidade de elementos da classe negativa é muito pequena para representar um resultado fidedigno. Tal resultado acaba não sendo uma surpresa, visto que apenas 25% do dataset entra para uma das classes nesta configuração. Se elevarmos o critério de boa avaliação para 4.6, obtendo assim 915 jogos com “avaliação baixa” e 1085 com “avaliação alta”, as médias de acurácia, precisão e recall alteram para 67.97%, 68.39% e 76.50% respectivamente. Com 4.7 estrelas como critério, temos 73.69%, 67.95% e 35.60% respectivamente, e os valores de verdadeiros negativos passam a ser bem maiores que o de verdadeiros positivos (a situação se inverte). Abaixo é mostrada um gráfico comparativo das medidas de performance com base nos diferentes pontos de corte utilizados (utilizando algoritmo SVM).

A graph of a graph with green and orange lines

Description automatically generated

Utilizando *Histogram Gradient Boosting* para um limiar de 4.5 estrelas, obtem-se acurácia, precisão e recall de 78.82%, 80.51% e 93.35% respectivamente, e com limiar de 4.6 tem-se 69.14%, 71.24% e 74.25% respectivamente, produzindo um resultado semelhante ao SVM. Abaixo pode-se ver um gráfico comparativo utilizando os dois diferentes pontos de corte para o algoritmo de HGB.

**A graph of a breakpoint

Description automatically generated with medium confidence**

Para a propriedade “highEngagement”, representando se um jogo teve quantidade significativa de avaliações de usuários, os modelos preditos utiliando *Support Vector Machines* atingiram uma média de valores de acurácia, precisão e recall de 71.16%, 70.22% e 66.95% respectivamente, utilizando 30.000 como critério para separação de classes, possuindo assim um melhor desempenho comparado ao modelo predito sob o atributo “goodRating”. Além disso, se alterarmos o critério para 20.000 ou 50.000, temos como acurácia, precisão e recall 71.92%, 72.64% e 79.64%, e 73.28%, 70,71% e 52,09% respectivamente. Abaixo está o gráfico comparativo relacionando os diferentes pontos de corte para o algoritmo de SVM.

A graph of a graph with green and orange lines

Description automatically generated

Utilizando o método do *Histogram Gradient Boosting*, obtemos acurácia, precisão e recall 74.07%, 73.16% e 71.35% respectivamente, tendo critério de divisão de 30.000. Para um critério de divisão de 20.000, temos 72.87%, 74.53% e 78.92% respectivamente. Abaixo é exibido um gráfico comparativo das medidas para o algoritmo de HGB.

A graph with different colored lines

Description automatically generated

Comparando com SVM, notamos uma semelhança entre os resultados, com HGB sendo somente um pouco melhor. Tais resultados sugerem uma relação significativa entre a descrição de um jogo e a quantidade de engajamento que ela atrai.

Analisando as tabelas TF-IDF resultantes do grupo de classes de alto engajamento e baixo engajamento, notou-se uma maior incidência de palavras como “slot”, “casino” e “battle” dentre as 30 mais frequentes, sugerindo temas que atraem mais usuários. Além disso, caracteres especiais utilizados para um formato de tópicos, como pontos de meia altura, também aparecem dentre as 10 palavras com mais escore. Por fim, palavras relacionadas a assinaturas, contas e períodos de teste também estão entre palavras que atraem mais atenção, sugerindo um maior engajamento quando jogos acrescentam detalhes sobre suas assinaturas e formas de cobrança já na descrição dos mesmos. Abaixo está a lista completa das palavras (radicais) com maior escore na tabela TF-IDF.

A graph of a number of colored bars

Description automatically generated with medium confidence

# DISCUSSÃO

Após analisar os resultados, notou-se uma relação significativa entre a descrição dos jogos e a quantidade de avaliações feitas, podendo-se atrair mais usuários com algumas palavras-chave específicas e com informações mais precisas, especialmente no quesito assinaturas in-game.

A fim de obter uma melhor avaliação média, no entanto, as análises feitas não foram eficazes o suficiente, sendo a maior limitação a quantidade de dados disponíveis no dataset. Dos 2000 jogos, 1879 já possuiam uma avaliação maior que 4 estrelas (o que normalmente consideramos como uma boa avaliação). Além disso, esta também foi uma limitação que poderia impedir o modelo relacionado à alto engajamento de obter uma melhor performance.

Finalmente, para o dataset e parâmetros utilizados, o algoritmo de *Histogram Gradient Boosting* apresentou desempenho muito semelhante ao do *Support Vector Machines*, obtendo resultados apenas um pouco melhores.

# CONCLUSÃO E TRABALHOS FUTUROS

Para trabalhos futuros, uma possibilidade seria aplicar esta mesma análise e testes em um dataset maior, coletando mais informações na loja para obter resultados mais fidedignos, inclusive analisando a relação entre descrição de um jogo e avaliação média.

Outro trabalho interessante pode também envolver buscar uma correlação entre capturas de telas de um jogo e suas avaliações, eventualmente classificando o tipo de capturas de tela, se possuem mais texto ou se são mais visuais, quais as cores utilizadas, etc.

Por fim, um objetivo inicial que o autor possuia mas que também não se tornou possível devido ao dataset, seria buscar uma relação entre o custo de instalação dos jogos e suas avaliações, junto à descrição. Com este dataset isto não se tornou possível devido à apenas um jogo possuir um custo de instalação diferente de zero.

# Referências

OMAR, Aisulu, **Is it possible to predict rating of google play store apps based on the given information?,** 2019. Disponível em: https://medium.com/analytics-vidhya/is-it-possible-to-predict-rating-of-google-play-store-apps-based-on-the-given-information-da9a44a3ac1e

DAIMI, Kevin, HAZZAZI, Noha. **Using Apple Store Dataset to Predict User Rating of Mobile Applications**. International Conference on Data Science, 2019

KARUNARATHNA, Kanchana. **GameSphere: 2000 App Store Insights & Ratings.** Unlock the Secrets of Player Preferences and Game Trends. 2024. Disponível em: https://www.kaggle.com/datasets/kanchana1990/gamesphere-2000-app-store-insights-and-ratings/data

APICHARDSILKIJ, Nattapoj. **Basic Comparison Between Random Forest, SVM, and XGBoost**. 2024. Disponível em: https://medium.com/@ap.nattapoj\_st/basic-comparison-between-randomforest-svm-and-xgboost-0e5862871175

Nayebi, M., Cho, H. & Ruhe, G. **App store mining is not enough for app improvement**. Empir Software Eng 23, 2764–2794 (2018). Disponível em: https://doi.org/10.1007/s10664-018-9601-1

Bin Fu, Jialiu Lin, Lei Li, Christos Faloutsos, Jason Hong, and Norman Sadeh. 2013. **Why people hate your app: making sense of user feedback in a mobile app store**. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13). Association for Computing Machinery, New York, NY, USA, 1276–1284. Disponível em: https://doi.org/10.1145/2487575.2488202

Islam, Mir. (2014). **Numeric Rating of Apps on Google Play Store by Sentiment Analysis on User Reviews**. Disponível em: https://doi.org/10.1109/ICEEICT.2014.6919058

Lin, D., Bezemer, CP., Zou, Y. et al. **An empirical study of game reviews on the Steam platform**. Empir Software Eng 24, 170–207 (2019). Disponível em: https://doi.org/10.1007/s10664-018-9627-4