Reconhecimento de Padrões

Análise de Jogos da Steam

Grupo:

- Guilherme Martiniano de Oliveira – 11215765
- Mateus Miquelino da Silva – 11208412
- Gustavo Fernandes Carneiro de Castro – 11369684

Professor:

Ricardo Zorzetto Nicoliello Vencio

Introdução

O *dataset* selecionado para a análise possuía diversas informações sobre jogos distribuídos pela plataforma *Steam*, a maior plataforma de distribuição online de jogos eletrônicos. Esse *dataset*, de nome *steam.csv,* foi retirado do site *Kaggle*, uma comunidade online de pesquisas em análise de dados. O *dataset* contém dados retirados da própria plataforma, e do site *SteamSpy*, que disponibiliza informações diversas sobre os jogos.

Objetivos

O objetivo da análise foi descobrir quais dos diferentes dados presentes no *steam.csv* eram relevantes para descobrir sua popularidade (número estimado de jogadores totais). Essa análise foi feita utilizando 3 métodos diferentes de análise supervisionada (KNN, com 3 e com 7 vizinhos mais próximos, e LDA). Além disso, os métodos foram comparados para descobrir o melhor método de análise possível para esse dataset.

Dataset

O dataset era inicialmente composto por 18 colunas totais e um total de 27075 linhas, cada uma com um jogo diferente.

As colunas eram:

appid	Identificador diferenciado para cada jogo.
name	Título do jogo.
release_date	Data de lançamento do jogo no formado AAAA-MM-DD.
english	Suporte de idioma: 1 se o jogo é em inglês.
developer	Nome(s) do(s) desenvolvedor(es). Ponto e vírgula para delimitar múltiplos desenvolvedores.
publisher	Nome(s) do(s) distribuidora(s). Ponto e vírgula para delimitar múltiplas distribuidoras.
platforms	Lista delimitada por pontos e vírgulas das plataformas suportadas.
required_age	Idade mínima requerida para jogar, de acordo com o padrão da PEGI UK. 0 é livre para todos os públicos.
categories	Lista delimitada por pontos e vírgulas das categorias dos jogos.
genres	Lista delimitada por pontos e vírgulas dos gêneros dos jogos.
steamspy_tags	Lista delimitada por pontos e vírgulas das *tags* mais populares dos jogos (similar aos gêneros, mas dados pela comunidade).
achievements	Número de conquistas dentro do jogo.
positive_ratings	Número de análises positivas, dados do *SteamSpy.*
negative_ratings	Número de análises negativas, dados do *SteamSpy.*
average_playtime	Média de tempo jogado (minutos), dados do *SteamSpy.*
median_playtime	Mediana de tempo jogado (minutos), dados do *SteamSpy.*
owners	Número estimado de donos. Contém o limite superior e inferior.
price	Preço do jogo na época, em libras esterlinas.

Para tratar o dataset, nós analisamos cada uma das colunas para ver se havia discrepância. Após verificar uma a uma utilizando *summary(steam)*, nós verificamos q as

colunas *average_playtime* e *median_playtime* tinham: os valores de tempo médio de jogo iguais aos da mediana dos tempos de jogo; e/ou os valores de tempo médio de jogo maiores que os valores do jogo mais jogado da *Steam*, o *Dota 2*. Após um simples loop, esses dados foram retirados, deixando o dataset com 27059 linhas (jogos).

Além disso, *steam.csv* possuía diversas colunas com dados não numéricos e em muita quantidade, como gênero do jogo e plataformas de lançamento. Esses dados eram do tipo "*string*", e separados por ponto e vírgula, como por exemplo: "Action; Adventure". Para separá-los, foram criadas matrizes *nrow(steam)x1*, com informações binárias (0 e 1, "tem" e "não tem", respectivamente) sobre a existência desses dados nas colunas do jogo, pela utilização da função *grepl("dado", steam[jogo, "coluna"])* para separar as "*strings*".

Após implementar os tratamentos de dado acima, nas colunas que foram selecionadas como relevantes, *steam.csv* ficou com um total de 27059 linhas (jogos) e 27 colunas (dados), nesta mesma ordem:

english	Dado binário. Possui, ou não, suporte de idioma (inglês).
required_age	Idade mínima requerida para jogar, de acordo com o padrão da PEGI UK.
achievements	Número de conquistas dentro do jogo.
positive_ratings	Número de análises positivas, dados do *SteamSpy.*
negative_ratings	Número de análises negativas, dados do *SteamSpy.*
average_playtime	Média de tempo jogado (minutos), dados do *SteamSpy.*
median_playtime	Mediana de tempo jogado (minutos), dados do *SteamSpy.*
price	Preço do jogo na época, em libras esterlinas.
windows	Dado binário. Possui, ou não, suporte para *Windows*.
linux	Dado binário. Possui, ou não, suporte para *Linux*.
mac	Dado binário. Possui, ou não, suporte para *Mac*.
free_to_play	Dado binário. É, ou não, de graça para instalar e jogar.
action	Dado binário. É, ou não, um jogo de ação.
indie	Dado binário. É, ou não, um jogo indie (criado por uma pequena empresa).
adventure	Dado binário. É, ou não, um jogo de aventura.
strategy	Dado binário. É, ou não, um jogo de estratégia.
rpg	Dado binário. É, ou não, um jogo de RPG.
racing	Dado binário. É, ou não, um jogo de corrida.
simulation	É, ou não, um jogo de simulação. Dado binário.
casual	Dado binário. É, ou não, um jogo casual.
sports	Dado binário. É, ou não, um jogo de esporte.
violent	Dado binário. É, ou não, um jogo violento.
nudity	Dado binário. Possui, ou não, nudez.
single_player	Dado binário. É, ou não, um jogo violento.
multiplayer	Dado binário. Pode, ou não, ser jogado sozinho.
steam_cards	Dado binário. Pode, ou não, ser jogado por múltiplos jogadores.
created_by_valve	Dado binário. É, ou não, um jogo criado pela própria *Steam* (a *Valve* é a criadora e distribuidora de jogos cuja *Steam* é dona).

Como os rótulos (labels) para a nossa análise, foi escolhida a coluna *owners*, que resume bem a popularidade de um jogo: quanto mais popular, mais pessoas tem o jogo.

Métodos

Depois de ajeitar o dataset, deu-se espaço às análises. Primeiramente, foi escolhida uma *seed* para os testes, *set.seed(6969).* Depois, foi estabelecido um grupo de *treino* com 80% dos jogos (escolhidos arbitrariamente pela função *sample()*). A comparação entre resultado das análises com os 20% restantes devolve a quantidade de acertos e erros e a precisão da análise, através da função da matriz de confusão, *confusionMatrix(resultado, labels[-treino],* retirada da biblioteca *caret)*.

Essa análise foi feita para o KNN, com os 3 e os 7 vizinhos mais próximos(KNN(3) e KNN(7)), e com o LDA. A quantidade de vizinhos foi escolhida propositalmente: 3 pois é a quantidade mais usual, e 7 pois foi a com a melhor precisão.

Primeiro, foi feita a análise dos KNNs e do LDA para o *treino* usual e anotada a precisão. Depois, 27 análises novas foram feitas para o KNN(3), o KNN(7) e para o LDA: a cada iteração, uma versão do *treino* sem uma coluna diferente era analisada. As precisões de cada iteração foram anotadas e comparadas.

Depois de comparadas, ao analisar que dado diminuía ou aumentava a precisão após ser retirado, foram descobertos os melhores dados (colunas) para cada método. Os *sub-treinos* de *treino* que possuíam apenas as melhores colunas foram chamados de *perfect_Method_* (*perfectKNN3, perfectKNN7* e *perfectLDA).*

A última análise foi feita após juntar os *perfect_Method_,* sendo chamados de *superPerfect_Method_*. Os dois *superPerfect_Method_* criados foram o *superPerfectKNN,* criado da junção do *perfectKNN3* e do *perfectKNN7; e o superPerfectAll*, criado da junção de todos os métodos de análise. A junção foi feita mantendo apenas os dados (colunas) presentes em todos os *perfect_Method_* que compunham o *superPerfect_Method_*.

Dados

Os dados retirados das análises acima seguem abaixo:

KNN(3)

Número da coluna retirada	Nome da coluna retirada	Precisão	Comparação	Importante ou Problemático
Original	0	0,6833		
-1	*english*	0,6833	Igual	Irrelevante
-2	*required_age*	0,6807	Diminuiu	Importante
-3	*achievements*	0,6853	Aumentou	Problema
-4	*positive_ratings*	0,6787	Diminuiu	Importante
-5	*negative_ratings*	0,6779	Diminuiu	Importante
-6	*average_playtime*	0,6800	Diminuiu	Importante
-7	*median_playtime*	0,6726	Diminuiu	Importante
-8	*price*	0,6964	Aumentou	Problema
-9	*windows*	0,6857	Aumentou	Problema
-10	*linux*	0,6848	Aumentou	Problema
-11	*mac*	0,6842	Aumentou	Problema
-12	*free_to_play*	0,6772	Diminuiu	Importante
-13	*action*	0,6805	Diminuiu	Importante
-14	*indie*	0,6844	Aumentou	Problema
-15	*adventure*	0,6839	Aumentou	Problema
-16	*strategy*	0,6864	Aumentou	Problema
-17	*rpg*	0,6874	Aumentou	Problema
-18	*racing*	0,6861	Aumentou	Problema
-19	*simulation*	0,6807	Diminuiu	Importante
-20	*casual*	0,6892	Aumentou	Problema
-21	*sports*	0,6855	Aumentou	Problema
-22	*violent*	0,6866	Aumentou	Problema
-23	*nudity*	0,6813	Diminuiu	Importante
-24	*single_player*	0,6844	Aumentou	Problema
-25	*multiplayer*	0,6907	Aumentou	Problema
-26	*steam_cards*	0,6888	Aumentou	Problema
-27	*created_by_valve*	0,6848	Aumentou	Problema

Nome do sub-treino	Número das colunas que se mantiveram	Precisão
Perfect KNN(3)	2, 4, 5, 6, 7, 12, 13, 19, 23	0,7350
Perfect KNN(7)	1, 4, 5, 6, 7, 9, 10, 11, 12, 13, 16, 17, 20, 21, 22, 23, 27	0,7164
Perfect LDA	1, 3, 4, 5, 6, 7, 8, 10, 11, 18, 20	0,7199
Super Perfect KNN	4, 5, 6, 7, 12, 23	0,7385
Super Perfect All	4, 5, 6, 7, 23	0,7234

KNN(7)

Número da coluna retirada	Nome da coluna retirada	Precisão	Comparação	Importante ou Problemático
Original	0	0,6949		
-1	*english*	0,6888	Diminuiu	Importante
-2	*required_age*	0,6962	Aumentou	Problema
-3	*achievements*	0,6977	Aumentou	Problema
-4	*positive_ratings*	0,6918	Diminuiu	Importante
-5	*negative_ratings*	0,6935	Diminuiu	Importante
-6	*average_playtime*	0,6949	Igual	Problema
-7	*median_playtime*	0,6912	Diminuiu	Importante
-8	*price*	0,6986	Aumentou	Problema
-9	*windows*	0,6929	Diminuiu	Importante
-10	*linux*	0,6942	Diminuiu	Importante
-11	*mac*	0,6938	Diminuiu	Importante
-12	*free_to_play*	0,6911	Diminuiu	Importante
-13	*action*	0,6992	Aumentou	Problema
-14	*indie*	0,6990	Aumentou	Problema
-15	*adventure*	0,6968	Aumentou	Problema
-16	*strategy*	0,6936	Diminuiu	Importante
-17	*rpg*	0,6944	Diminuiu	Importante
-18	*racing*	0,6960	Aumentou	Problema
-19	*simulation*	0,6959	Aumentou	Problema
-20	*casual*	0,6936	Diminuiu	Importante
-21	*sports*	0,6914	Diminuiu	Importante
-22	*violent*	0,6921	Diminuiu	Importante
-23	*nudity*	0,6949	Igual	Irrelevante
-24	*single_player*	0,6962	Aumentou	Problema
-25	*multiplayer*	0,6979	Aumentou	Problema
-26	*steam_cards*	0,6986	Aumentou	Problema
-27	*created_by_valve*	0,6914	Diminuiu	Importante

Nome do sub-treino	Número das colunas que se mantiveram	Precisão
Perfect KNN(3)	2, 4, 5, 6, 7, 12, 13, 19, 23	0,7428
Perfect KNN(7)	1, 4, 5, 6, 7, 9, 10, 11, 12, 13, 16, 17, 20, 21, 22, 23, 27	0,7413
Perfect LDA	1, 3, 4, 5, 6, 7, 8, 10, 11, 18, 20	0,7313
Super Perfect KNN	4, 5, 6, 7, 12, 23	0,7493
Super Perfect All	4, 5, 6, 7, 23	0,7408

LDA

Número da coluna retirada	Nome da coluna retirada	Precisão	Comparação	Importante ou Problemático
Original	0	0,6883		
-1	*english*	0, 6881	Diminuiu	Importante
-2	*required_age*	0, 6896	Aumentou	Problema
-3	*achievements*	0, 6883	Igual	Irrelevante
-4	*positive_ratings*	0, 6874	Diminuiu	Importante
-5	*negative_ratings*	0, 6855	Diminuiu	Importante
-6	*average_playtime*	0, 6877	Diminuiu	Importante
-7	*median_playtime*	0, 6866	Diminuiu	Importante
-8	*price*	0, 6883	Igual	Irrelevante
-9	*windows*	0, 6885	Aumentou	Problema
-10	*linux*	0, 6875	Diminuiu	Importante
-11	*mac*	0, 6864	Diminuiu	Importante
-12	*free_to_play*	0, 6896	Aumentou	Problema
-13	*action*	0, 6888	Aumentou	Problema
-14	*indie*	0, 6916	Aumentou	Problema
-15	*adventure*	0, 6888	Aumentou	Problema
-16	*strategy*	0, 6887	Aumentou	Problema
-17	*rpg*	0, 6885	Aumentou	Problema
-18	*racing*	0, 6883	Igual	Irrelevante
-19	*simulation*	0, 6888	Aumentou	Problema
-20	*casual*	0, 6881	Diminuiu	Importante
-21	*sports*	0, 6877	Aumentou	Problema
-22	*violent*	0, 6881	Aumentou	Problema
-23	*nudity*	0, 6885	Aumentou	Problema
-24	*single_player*	0, 6885	Aumentou	Problema
-25	*multiplayer*	0, 6874	Aumentou	Problema
-26	*steam_cards*	0, 6896	Aumentou	Problema
-27	*created_by_valve*	0, 6894	Aumentou	Problema

Nome do sub-treino	Número das colunas que se mantiveram	Precisão
Perfect KNN(3)	2, 4, 5, 6, 7, 12, 13, 19, 23	0,6864
Perfect KNN(7)	1, 4, 5, 6, 7, 9, 10, 11, 12, 13, 16, 17, 20, 21, 22, 23, 27	0,6912
Perfect LDA	1, 3, 4, 5, 6, 7, 8, 10, 11, 18, 20	0,6911
Super Perfect KNN	4, 5, 6, 7, 12, 23	0,6909
Super Perfect All	4, 5, 6, 7, 23	0,6912

Conclusão

Após analisar os dados, é possível concluir que o melhor método, encontrado pelas nossas análises, para definir quão popular é um jogo é o KNN(7). Além disso, os dados (colunas) que fazem um jogo se aproximar mais dos jogos com mais usuários são: *positive_ratings, negative_ratings, average_playtime, median_playtime, free_to_play* e *nudity,* retirado do Super Perfect KNN, o *sub-treino* com maior precisão no KNN(7).

Sem levar em conta o *free_to_play* e o *nudity*, todos os outros dados do SuperPerfectKNN são gerados quando o jogo já está em mercado. Isso nos leva a concluir que não há preferência de gênero dos usuários da *Steam* no geral, e que a qualidade do jogo é mais importante que seu estilo, gênero, local de lançamento e coletáveis.

Além disso, o fato de o jogo ser *free_to_play* é muito relevante, pois qualquer um pode adicionar um jogo grátis na sua biblioteca de jogos, significando que muito mais gente possui o jogo.

Referências

Dataset:
- https://www.kaggle.com/nikdavis/steam-store-games/

Auxílio no tratamento de dados:
- https://steamspy.com/
- https://store.steampowered.com/

Auxílio com a linguagem:
- https://www.rdocumentation.org/
- https://www.r-bloggers.com/installing-r-packages/