

Análise Descritiva



Grupo F:

Ana Pontes¹

Fábio Nunes²

Marta Pinhal³

¹up202005126@up.pt

²up202008145@up.pt

³up2020094372@up.pt

Mestrado em Ciência da Informação

Análise e Visualização de dados

2023/2024

Sumário

1	Introdução.....	5
2	Metodologia Utilizada.....	6
3	Compreensão do Negócio.....	7
3.1	Identificação do Problema	7
4	Identificação e Compreensão dos Dados.....	8
4.1	Descrição dos Atributos	8
4.2	Correlação entre Atributos	11
4.3	Qualidade dos Dados	13
5	Preparação dos Dados, Modelação e Avaliação	14
5.1	Preparação de dados	14
5.2	Modelação	19
5.2.1	Árvore de Decisão	19
5.3	Avaliação	22
6	Conclusão.....	25
	Referências	26
	Anexos.....	27
	Anexo 1	27

Índice de Figuras

Figura 1 - Metodologia CRISP-DM	6
Figura 2 - Apresentação geral do dataset (parte 1)	8
Figura 3 - Apresentação geral do dataset (parte 2)	9
Figura 4 - Processo de Correlação no RapidMiner.....	11
Figura 5 - Resultado do Processo de Correlação	12
Figura 6 - Design da junção da tabela “games” e “recommendations”	14
Figura 7 - Lista de parâmetros dos papéis (“games” e “recommendations”)	15
Figura 8 - Lista de parâmetros dos atributos chave (“games” e “recommendations”).....	15
Figura 9 - Design da junção da tabela “games, “recommendations” e “users”	16
Figura 10 - Lista de parâmetros dos papéis (“games, “recommendations” e “users”)	16
Figura 11 - Lista de parâmetros dos atributos chave (“games, “recommendations” e “users”)	17
Figura 12 - Design da organização do título	17
Figura 13 - Rating	18
Figura 14 - Conversão de Nominal para Numérico.....	18
Figura 15 - Criação do Modelo "Decision Tree"	19
Figura 16 - Atribuição do Label aos atributos "positive_ratio" e “title”	20
Figura 17 - Definição dos Parâmetros do Modelo	20
Figura 18 - Resultado da Árvore de Decisão	21
Figura 19 - Estatísticas relativas aos Atributos	27

Índice de Tabelas

Tabela 1 - Descrição dos Atributos	10
Tabela 2 - Interpretação da situação conforme o Tipo de Rating	23

Resumo: O foco do presente trabalho foi o dataset "Game Recommendations on Steam", composto por games.csv, users.csv e recommendations.csv. Tendo depois passado para a criação de uma empresa fictícia "Game+", que foi criada com o objetivo de oferecer recomendações de jogos dependendo do perfil de cada utilizador. Após uma análise descritiva e compreensão dos dados, realizou-se a seleção e preparação dos dados para modelagem, tendo de seguida, passado para a fase de modelagem envolveu a escolha e justificação de técnicas, a construção e avaliação de modelos, seguidos por uma revisão para determinar os próximos passos. Finalmente, com base nos resultados e análises, conclusões serão elaboradas para relacionar o tema com a sua aplicação prática no mundo real.

Palavras-Chave: Análise de Dados, Análise Descritiva, RapidMiner, Árvore de Decisão, Metodologia CRISP-DM, Recomendação de jogos, Atributos.

Summary: The focus of this work was the "Game Recommendations on Steam" dataset, made up of games.csv, users.csv and recommendations.csv. It then moved on to the creation of a fictitious company "Game+", which was created with the aim of offering game recommendations depending on each user's profile. After a descriptive analysis and understanding of the data, the data was selected and prepared for modelling, and then the modelling phase involved choosing and justifying techniques, building and evaluating models, followed by a review to determine the next steps. Finally, based on the results and analyses, conclusions will be drawn to relate the topic to its practical application in the real world.

Keywords: Data Analysis, Descriptive Analysis, RapidMiner, Decision Tree, CRISP-DM Methodology, Game Recommendation, Attributes.

1 Introdução

No âmbito da unidade curricular Análise e Visualização de Dados, lecionada no primeiro semestre do primeiro ano do Mestrado em Ciência da Informação, realizou-se o presente trabalho, com o objetivo de aplicar o conhecimento adquirido numa situação real através da metodologia CRISP-DM.

Primeiramente foi necessário escolher um Data Set que permitisse uma análise descritiva e posteriormente preditiva. De seguida os dados escolhidos foram compreendidos, descritos, explorados, alterados e posteriormente avaliados.

Assim sendo, o dataset escolhido foi o “**Game Recommendations on Steam**”, que compreende três entidades principais - games.csv (uma tabela com dados sobre jogos ou *add-ons*, contendo informações como classificações, preços em dólares americanos, data de lançamento, entre outros. Um arquivo de metadados complementa esses dados com detalhes não tabulares, como descrições e etiquetas), users.csv (uma tabela contendo informações públicas sobre perfis de utilizadores, incluindo o número de produtos adquiridos e críticas publicadas) e recommendations.csv (uma tabela que registra avaliações de utilizadores, indicando se recomendam ou não um produto. Esta tabela representa uma relação muitos-muitos entre jogos e utilizadores). Este conjunto de dados refere-se a jogos, utilizadores e críticas, utilizado na criação de sistemas de recomendação. A loja Steam, plataforma líder para a aquisição e transferência de jogos, DLCs e conteúdo relacionado, foi então o ponto de partida para o desenvolvimento do dataset.

Depois de escolhido o domínio e o dataset que será trabalhado, passamos à fase de construção e desenvolvimento de uma empresa fictícia. Assim, criamos a “Game +”, uma plataforma impulsionada pela paixão por jogos, que visa preencher a lacuna entre jogadores e a vasta gama de jogos disponíveis. A missão da empresa é então oferecer uma jornada personalizada e envolvente no universo dos jogos, fornecendo recomendações precisas e relevantes que se alinham às preferências individuais do utilizador.

Após a seleção do conjunto de dados, conduziu-se uma análise descritiva, compreendendo inicialmente os objetivos do negócio. A partir dessa compreensão, formulou-se um problema que seria abordado por meio de um plano estruturado. Em seguida, procedeu-se à interpretação dos dados, descrevendo os atributos e avaliando a sua qualidade. Na etapa subsequente, foram escolhidos os dados considerados essenciais para resolver o problema identificado, realizando-se a limpeza, construção, integração e formatação do conjunto de dados para a preparação da fase de modelagem.

Na fase de modelagem, foi necessário decidir sobre a técnica a ser utilizada, justificando a escolha. Após a construção do modelo, os resultados foram avaliados para atender o problema identificado. Este processo passou por uma revisão para determinar os próximos passos do projeto.

No final, com base em todos os resultados alcançados neste projeto e na análise descritiva realizada, serão tiradas conclusões para estabelecer a conexão deste tema com sua aplicação prática num contexto real.

2 Metodologia Utilizada

A metodologia adotada neste trabalho prático é baseada na metodologia CRISP-DM. Esta metodologia é estruturada em fases distintas, tais como: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e desenvolvimento, as quais serão detalhadamente abordadas de seguida.

A fase inicial, de compreensão do negócio, concentra-se na compreensão dos objetivos do negócio, realizando uma avaliação da situação para, a partir disso, definir um problema de análise de dados e elaborar um plano para superar esse desafio.

O próximo passo consiste na compreensão dos dados, onde ocorre a recolha do conjunto de dados a ser utilizado no projeto. Após a recolha, é fundamental descrever esses dados, examinando não apenas o volume, mas também as principais propriedades. Além disso, esta fase abrange a identificação da acessibilidade e disponibilidade dos atributos do conjunto de dados, descrevendo o tipo de atributos, intervalos e correlações entre os dados. Esses procedimentos visam verificar a qualidade dos dados recolhidos.

A terceira fase é dedicada à preparação dos dados, na qual, por meio do conjunto de dados, são selecionados e preparados os dados brutos iniciais que vão ser utilizados. Essa seleção inclui a correção e limpeza dos dados incorretos ou desnecessários para a resolução do problema em questão. Posteriormente, passamos à fase de modelação dos dados, na qual são selecionadas e aplicadas técnicas de modelação para enfrentar os problemas identificados. Em alguns casos, várias técnicas podem ser aplicáveis para resolver o mesmo tipo de problema, tornando inevitável a análise de todas as opções para determinar a mais adequada à situação em questão. Frequentemente, é necessário retornar à fase de preparação dos dados para viabilizar a utilização da técnica escolhida.

A fase subsequente é a avaliação dos dados gerados pelo modelo utilizado. Nessa avaliação, a ênfase recai sobre a qualidade por meio da análise dos dados, avaliando o modelo e, com base nessa análise, decidindo como utilizar os resultados obtidos para enfrentar os problemas identificados na fase de compreensão do negócio. Em seguida, realiza-se uma revisão minuciosa para verificar se o modelo atinge os objetivos e se há algum problema não identificado. Dependendo da revisão, são determinados os próximos passos, como a análise do potencial de cada resultado e a identificação de melhorias que podem ser implementadas no processo atual.

Como fase final, temos o desenvolvimento, cujo objetivo é aplicar o conhecimento adquirido na análise dos dados e compreender como esses modelos podem ser aplicados em contexto real.

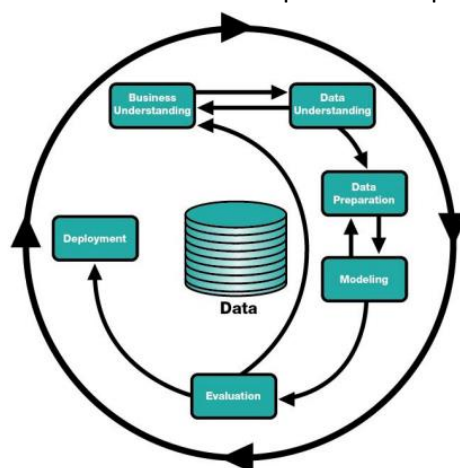


Figura 1 - Metodologia CRISP-DM

3 Compreensão do Negócio

Na fase da Compreensão do Negócio da metodologia *CRISP-DM*, o foco está na compreensão profunda dos objetivos e necessidades da organização. Durante esta etapa, realiza-se uma análise minuciosa da situação atual do negócio, avaliando os seus objetivos, desafios e contextos específicos. O objetivo é identificar claramente os problemas que podem ser abordados por meio da análise de dados.

Nesta etapa, o procurado é compreender a dinâmica operacional, os objetivos estratégicos e as questões-chave para o contexto da organização. Isso inclui a definição de um problema de análise de dados que seja relevante se dar resposta através do dataset para os objetivos globais da organização. No final da etapa da compreensão do negócio, espera-se ter uma visão clara do contexto organizacional e dos desafios que a análise de dados pode ajudar a resolver.

3.1 Identificação do Problema

O negócio a analisar diz respeito a uma empresa de plataforma online de jogos que detém um dataset cujos dados servem para uma boa e mais correta recomendação de jogos dependendo do perfil do utilizador. Assim, as recomendações de utilizadores, informações sobre jogos da Loja Steam, levou-nos à identificação de um potencial problema a: "Previsão de Popularidade Futura de Jogos".

Dado o conjunto de dados que inclui avaliações de utilizadores e informações sobre jogos, este problema preditivo possibilitava a previsão da popularidade futura de jogos na plataforma Steam. Isso envolveria a construção de um modelo que leva em consideração as características dos jogos, o histórico de avaliações, o feedback dos utilizadores e outras informações relevantes para prever quais jogos têm maior probabilidade de serem jogados, baseados nas avaliações positivas, no futuro. Essa previsão poderia ajudar a plataforma a destacar e promover jogos que provavelmente atrairão mais utilizadores, melhorando assim a eficácia do sistema de recomendação.

O objetivo é então, desenvolver um modelo preditivo capaz de analisar diversas características dos jogos (como preço, data de lançamento, título, plataforma onde é suportado, entre outros) e do comportamento dos utilizadores (avaliações, padrões de compra, etc.) para antecipar quais jogos terão maior adesão e popularidade entre os utilizadores no futuro.

A metodologia passaria pela:

- Identificação e compreensão de dados: Compreender as características e estruturas das tabelas nos conjuntos de dados.
- Agregação de Dados: Utilização de dados históricos da Steam, incluindo avaliações dos utilizadores, informações sobre jogos e perfis de utilizadores.
- Pré-processamento: Limpar e preparar os dados, identificando características relevantes para a previsão de popularidade.
- Modelagem Preditiva: Escolher algoritmos de *machine learning* adequados para criar um modelo preditivo com base nos dados preparados.

4 Identificação e Compreensão dos Dados

Nesta fase tornou-se essencial proceder a uma exploração mais detalhada dos dados para analisar.

O procedimento de compreensão de dados dividiu-se então em duas etapas: visualização e análise geral do dataset em excel e importação do dataset para o RapidMiner e exploração das estatísticas.

Sendo que tínhamos um elevado número de dados em cada uma das tabelas decidimos eliminar algumas linhas e seguirmos com 50873 linhas em cada uma das tabelas, sendo assim fizemos apenas uma análise de uma amostra do dataset inteiro por forma a ser possível analisarmos os dados nos computadores.

Ao fazer a junção das 3 tabelas (através do *inner join*): “games”, “Recommendations” e “users” conseguimos obter os seguintes dados para análise: 182 instâncias e 22 colunas, sendo que 19 destas correspondem a atributos do tipo: polinomial, número inteiro, data, real, e as outras 3 correspondem a diferentes ID’s.

4.1 Descrição dos Atributos

Antes da apresentação pormenorizada dos atributos, segue-se uma apresentação geral do dataset, assim sendo, na [Figura 2](#) e [Figura 3](#) é possível observar um excerto dos dados:

Open in

Turbo Prep

Auto Model

Filter (182 / 182 examples): all

Row No.	app_id	user_id	win	mac	linux	rating	is_recommended	title	date_release	positive_ratio	user_reviews	price_final	price_original	discount
1	346110	11457485	1	1	0	0	0	ARK: Survival Evolved	Aug 27, 2017 ...	83	495087	15	0	0
2	346110	6177428	1	1	0	0	0	ARK: Survival Evolved	Aug 27, 2017 ...	83	495087	15	0	0
3	346110	27882	1	1	0	0	1	ARK: Survival Evolved	Aug 27, 2017 ...	83	495087	15	0	0
4	1466860	5236314	1	0	0	0	0	Age of Empires IV: Anniversary Edition	Oct 28, 2021 ...	86	41462	40	0	0
5	270880	7977814	1	1	1	4	0	American Truck Simulator	Feb 2, 2016 1...	96	108202	20	0	0
6	270880	9120943	1	1	1	4	0	American Truck Simulator	Feb 2, 2016 1...	96	108202	20	0	0
7	1172470	5633784	1	0	0	0	0	Apex Legends	Nov 4, 2020 1...	80	713182	0	0	0
8	107410	5298042	1	1	0	0	0	Arma 3	Sep 12, 2013...	91	154094	30	0	0
9	107410	13480540	1	1	0	0	0	Arma 3	Sep 12, 2013...	91	154094	30	0	0
10	107410	5624389	1	1	0	0	0	Arma 3	Sep 12, 2013...	91	154094	30	0	0
11	244210	1383854	1	0	0	0	0	Assetto Corsa	Dec 19, 2014...	92	79527	20	0	0
12	244210	4317627	1	0	0	0	0	Assetto Corsa	Dec 19, 2014...	92	79527	20	0	0
13	244210	5983710	1	0	0	0	0	Assetto Corsa	Dec 19, 2014...	92	79527	20	0	0
14	371970	8778538	1	1	1	0	0	Barony	Jun 23, 2015 ...	92	3713	20	0	0
15	602960	6701099	1	1	1	0	0	Barotrauma	Mar 13, 2023 ...	93	35639	24	0	0
16	284160	10765119	1	0	0	4	0	BeamNG.drive	May 29, 2015 ...	97	178635	25	0	0
17	284160	8119841	1	0	0	4	0	BeamNG.drive	May 29, 2015 ...	97	178635	25	0	0
18	620980	11359796	1	0	0	4	0	Beat Saber	May 21, 2019 ...	95	63695	30	0	0
19	582660	9247869	1	0	0	3	0	Black Desert	May 24, 2017 ...	76	49539	10	0	0
20	397540	4826886	1	0	0	0	0	Borderlands 3	Mar 13, 2020 ...	85	95243	60	0	0
21	397540	11129719	1	0	0	0	0	Borderlands 3	Mar 13, 2020 ...	85	95243	60	0	0
22	397540	13182821	1	0	0	0	0	Borderlands 3	Mar 13, 2020 ...	85	95243	60	0	0
23	1938090	8960193	1	0	0	2	0	Call of Duty	Oct 27, 2022 ...	59	429206	0	0	0
24	255710	11379804	1	1	1	0	0	Cities: Skylines	Mar 10, 2015 ...	93	178458	30	0	0

Figura 2 - Apresentação geral do dataset (parte 1)

helpful	funny	date	hours	review_id	products	reviews
0	0	Jul 3, 2017 1...	113.100	2310	367	2
0	0	Jun 19, 2015 ...	34.700	46278	67	2
3	0	Sep 26, 2017...	832.100	47070	300	8
0	0	Nov 1, 2021 1...	31.600	32486	317	11
0	0	Jan 5, 2022 1...	84.500	14524	73	3
0	0	Oct 15, 2022 ...	199.900	31809	25	5
3	0	Jun 8, 2021 1...	570	47368	365	1
0	0	Oct 31, 2018 ...	481.500	20131	65	12
0	0	Jul 1, 2019 1...	316.800	26624	175	13
0	0	Jan 15, 2018 ...	19.300	43867	140	15
0	0	Nov 9, 2013 1...	14.600	30122	100	9
0	0	Nov 26, 2019 ...	98.100	34275	47	1
0	0	Aug 4, 2021 1...	424.200	43463	16	5
0	0	Nov 1, 2021 1...	29.100	46477	49	3
2	0	Jun 5, 2019 1...	54.900	2025	583	65
0	0	Jun 17, 2015 ...	400.400	9086	156	5
0	0	May 6, 2020 1...	751.100	18561	5	1
0	0	Jan 31, 2020 ...	48.100	10206	717	6
0	2	Jul 16, 2018 ...	0.400	43507	78	5
0	0	Mar 29, 2020 ...	28	22841	46	1
0	0	Mar 18, 2020 ...	52.100	24057	1344	5
0	0	Nov 25, 2020 ...	118.100	24131	493	10
0	0	Nov 11, 2022 ...	118.900	21133	83	1
0	0	Nov 24, 2017 ...	35.800	14054	949	15

Figura 3 - Apresentação geral do dataset (parte 2)

Tendo em conta os atributos e respetivos dados de cada atributo, segue-se uma tabela com a explicação dos atributos após exploração dos mesmos:

Atributo	Descrição	Formato
App_id	Identificação única do jogo	Integer
User_id	Identificação única do user	Integer
title	Título do jogo	Polynomial
Date_release	Data de lançamento do jogo	Date
win	Se tem suporte no Windows	Polynomial
mac	Se tem suporte no MacOS	Polynomial
linux	Se tem suporte no Linux	Polynomial
rating	Classificação do jogo	Polynomial
Positive_ratio	Rácio de avaliação positiva	Integer
User_reviews	Nº de avaliações dos utilizadores	Integer
Price_final	Preço em dólares americanos \$ calculado após o desconto	Real
Price_original	Preço em dólares americanos \$ antes do desconto	Real
Discount	Percentagem de desconto do jogo	Real
Steam_deck	Se tem suporte na plataforma da Steam	Polynomial
Helpful	Nº de utilizadores que consideraram o jogo útil	Integer
Funny	Nº de utilizadores que consideraram o jogo divertido	Integer
Date	Data de disponibilização do jogo na steam	Date
Is_recommended	Se o jogo é recomendado ou não	Polynomial
hours	Nº de horas contabilizadas	Real
Review_id	Identificação única da classificação	Integer
Products	Nº de jogos que o utilizador possui	Integer
Reviews	Quantidade de vezes que o user classificou um jogo	Integer

Tabela 1 - Descrição dos Atributos

Esta descrição de atributos que possibilita que se obtenha uma primeira noção dos atributos mais relevantes para a análise, assim como dos que não adicionam informação relevante.

Posto isto, retiramos uma série de imagens estatísticas por forma a ser possível observar estatísticas referentes a cada atributo, como se pode observar no **Anexo 1**.

Ao analisarmos os dados a partir do RapidMiner é possível verificar que os atributos presentes são do tipo “Número inteiro”, “Número Real”, “Polinomial” e “Data”.

Posteriormente é feita uma análise estatística apresentando os dados sobre o formato de gráfico de barras, para os dados polinomiais, é indicado qual o valor com menor e maior ocorrência e a contagem final da ocorrência de cada valor. No caso dos dados em número inteiro é indicado o número mínimo e máximo, a média dos valores e o desvio. Para a data é indicada a data mais recente e data mais longínqua, bem como a duração.

Concluimos assim com esta análise dos dados que quanto à sua qualidade que não se verifica a presença de dados em falta, dados inconsistentes, dados redundantes, “noisy data” e outliers.

4.2 Correlação entre Atributos

Para uma compreensão mais aprofundada dos dados, é essencial investigar se os atributos estão inter-relacionados e, se assim for, compreender a relevância dessa relação para a nossa análise. Nesse sentido, iniciamos importando o conjunto de dados para o RapidMiner e aplicando a matriz de correlação. Para otimizar a análise, ajustamos parâmetros, como a seleção exclusiva de atributos quantitativos e o cálculo da “squared correlation”, em vez de apresentar apenas correlações simples.

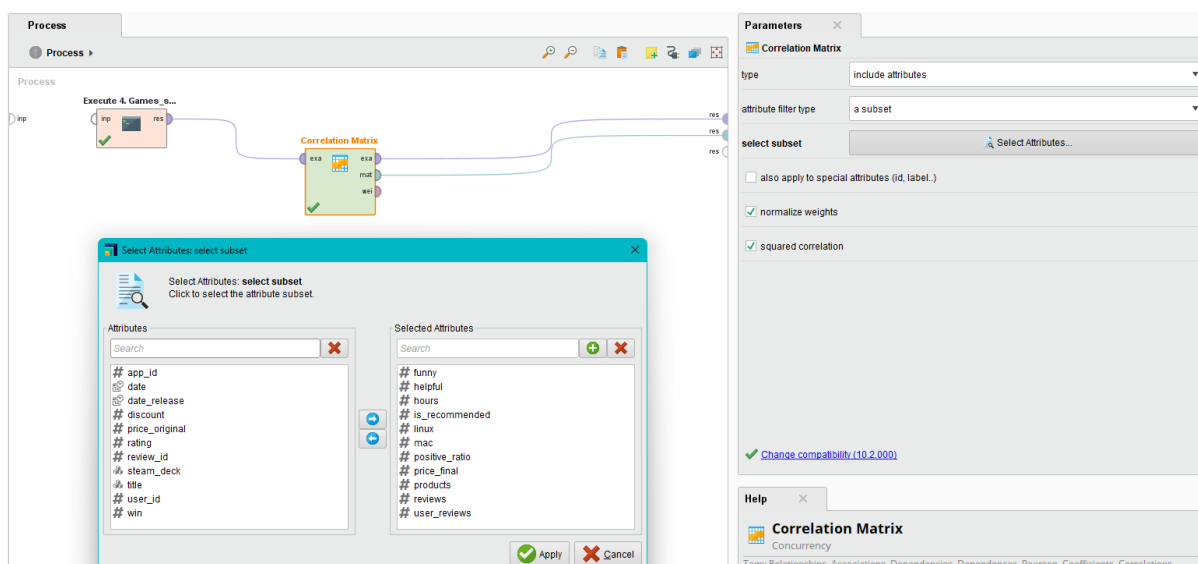


Figura 4 - Processo de Correlação no RapidMiner

Attributes	mac	linux	is_recommended	positive_ratio	user_reviews	price_final	helpful	funny	hours	products	reviews
mac	1	0.431	0.003	0.064	0.038	0.001	0.000	0.000	0.001	0.001	0.001
linux	0.431	1	0.003	0.029	0.078	0.000	0.004	0.000	0.005	0.000	0.002
is_recommended	0.003	0.003	1	0.044	0.001	0.002	0.011	0.001	0.015	0.009	0.000
positive_ratio	0.064	0.029	0.044	1	0.004	0.001	0.016	0.001	0.004	0.001	0.020
user_reviews	0.038	0.078	0.001	0.004	1	0.046	0.000	0.001	0.059	0.008	0.004
price_final	0.001	0.000	0.002	0.001	0.046	1	0.034	0.004	0.021	0.002	0.020
helpful	0.000	0.004	0.011	0.016	0.000	0.034	1	0.163	0.000	0.000	0.002
funny	0.000	0.000	0.001	0.001	0.001	0.004	0.163	1	0.003	0.000	0.001
hours	0.001	0.005	0.015	0.004	0.059	0.021	0.000	0.003	1	0.016	0.025
products	0.001	0.000	0.009	0.001	0.008	0.002	0.000	0.000	0.016	1	0.049
reviews	0.001	0.002	0.000	0.020	0.004	0.020	0.002	0.001	0.025	0.049	1

Figura 5 - Resultado do Processo de Correlação

Ao analisar a matriz de correlação é possível perceber sob o ponto de vista geral que os atributos não apresentam uma correlação entre si significativa.

A maior correlação que se pode verificar é entre "Mac" e "Linux", que é de 0,431, indica uma correlação positiva moderada entre a quantidade de sistemas operativos (Mac e Linux) compatíveis com os jogos. Existe ainda a relação entre o sistema operativo Windows (win) que não pode ser colocada na tabela de correlação, pois o mesmo têm uma taxa de compatibilidade com os jogos de 100%, sendo que este iria ter uma correlação de 1, podemos ainda assim concluir que o Windows é o sistema operativo mais usado pelos utilizadores que tem melhor compatibilidade com os jogos.

A segunda correlação mais alta que existe é entre "helpful" e "funny", com uma correlação de 0,163, sendo esta uma correlação positiva forte entre avaliações consideradas úteis e engraçadas.

As correlações positivas moderadas são as seguintes:

- "Linux" e "user_reviews" (0,078): Correlação positiva moderada entre a presença do sistema operativo Linux e o número de avaliações dos utilizadores. Verifica-se que alguns utilizadores recomendam o uso do sistema operativo Linux para um determinado jogo.
- "Products" e "reviews" (0,049): Correlação positiva moderada entre o número de produtos e o número de avaliações dos mesmos. Pode-se verificar que os jogos tem avaliações dos utilizadores.
- "price_final" e "helpful" (0,034): Correlação positiva moderada entre o preço final e se a avaliação foi útil. Verifica-se que o utilizador faz uma recomendação tendo em conta o jogo e o preço, sendo que estas demonstram-se úteis para os outros utilizadores que possam querer adquirir o jogo.
- "hours" e "reviews" (0,025): Correlação positiva moderada entre o número de horas jogadas e o número de classificações. Podemos verificar que os utilizadores ao jogarem um determinado número de horas fazem uma avaliação do jogo.
- "Is_recommended" e "hours" (0,015): Correlação positiva moderada entre a variável de recomendação e o número de horas jogadas. Quantas mais horas um jogador tiver no jogo irá fazer uma recomendação deste.

As correlações baixas e muito baixas que podemos encontrar foram as seguintes:

- "positive_ratio" e "user_reviews" (0,004): Correlação baixa entre o rácio de avaliação positiva e o número de avaliações dos utilizadores.
- "price_final" e "funny" (0,004): Correlação muito baixa entre o preço final e a variável de avaliações engraçadas.

- “Mac” e “Is_recommended” (0,003): Correlação muito baixa entre a presença de Mac e a variável recomendada. Podemos verificar que os utilizadores não recomendam usar o sistema operativo Mac para jogar.
- “price_final” e “Products” (0,002): Correlação muito baixa entre o preço final e o número de produtos.
- “positiveratio” e “price_final” (0,001): Correlação muito baixa entre o rácio de avaliação positiva e o preço final.

4.3 Qualidade dos Dados

A análise dos dados conduz-nos a um aspeto crucial: a sua qualidade, a qual desempenha um papel fundamental na possível influência sobre os resultados futuros. A qualidade dos dados revela-se na ausência significativa de elementos indesejáveis, tais como dados ruidosos, nulos, inconsistentes, redundantes, irregulares, valores em falta, *outliers*, entre outros.

Neste contexto, a nossa atenção concentrou-se predominantemente nestes aspetos, e é relevante salientar que não identificamos a presença dos referidos tipos de dados durante a nossa observação.

5 Preparação dos Dados, Modelação e Avaliação

5.1 Preparação de dados

A abordagem adotada para a preparação dos dados, conforme a Metodologia CRISP-DM aplicada neste relatório, incide sobre cinco pontos fundamentais: seleção, limpeza, construção, integração e formatação de dados. O propósito deste procedimento reside, primordialmente, na construção de um conjunto de dados final relevante a partir dos dados originais em estado bruto. Neste procedimento estão envolvidas tarefas como a escolha criteriosa dos dados que vão ser manipulados, a projeção das relações entre atributos para a resolução do problema apresentado na fase inicial, a preparação do conjunto de dados e a especificação do formato necessário para a análise subsequente.

Para fazer a junção das 3 tabelas foram realizados alguns processos que incluíram o dataset pretendido, o set role (para identificar o ID que era pretendido) e por fim o Join para agregar as tabelas tendo que na opção “key attributes” tivemos que identificar nos atributos chave da direita e da esquerda quais eram os pretendidos.

Posto isto realizou-se um processo inicial que iria juntar a tabela “games” com a tabela “recommendations”, para isso o set role de cada uma iria ser a “app_id”, pois é o ID comum dessas duas tabelas. Por fim foi então realizado um “join” para se juntar as duas tabelas sendo que queremos que este tenha em conta que para cada ID do jogo há uma recomendação diferente, posto isto é necessário identificar os atributos chave do lado direito e esquerdo por forma a nenhum ficar perdido.

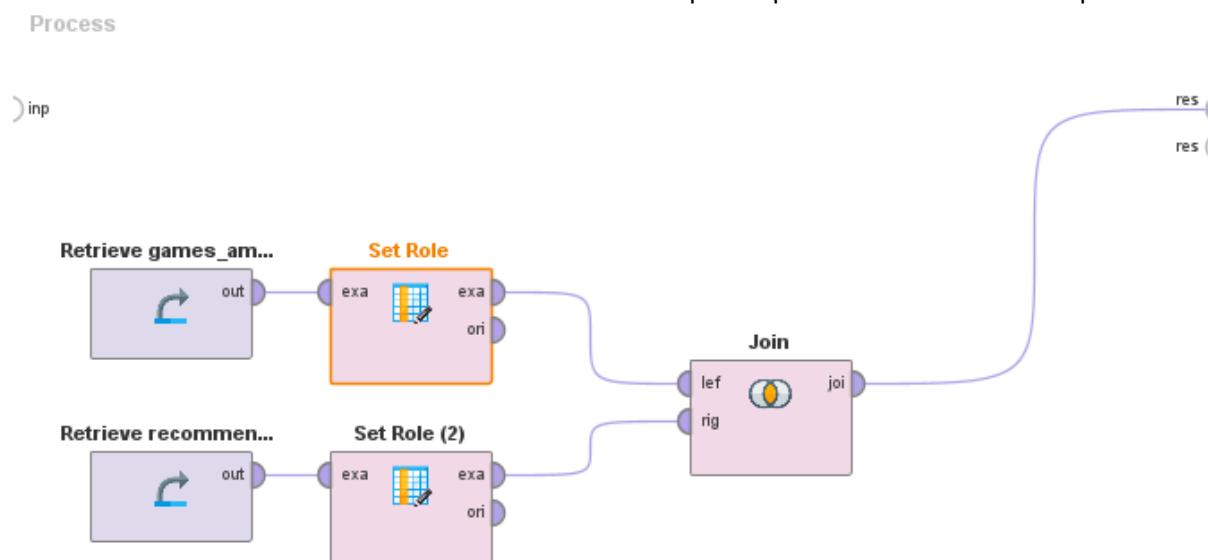
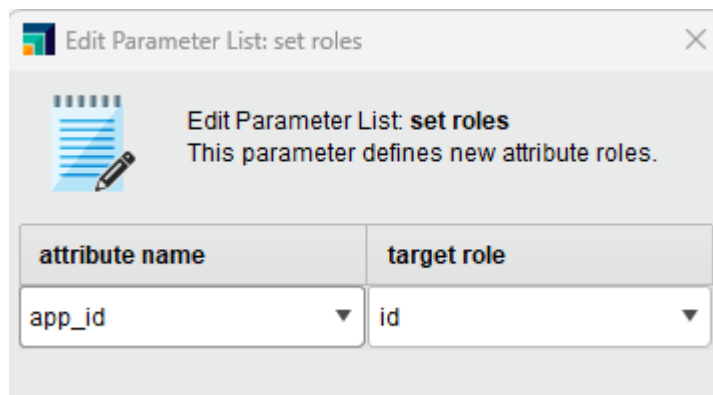


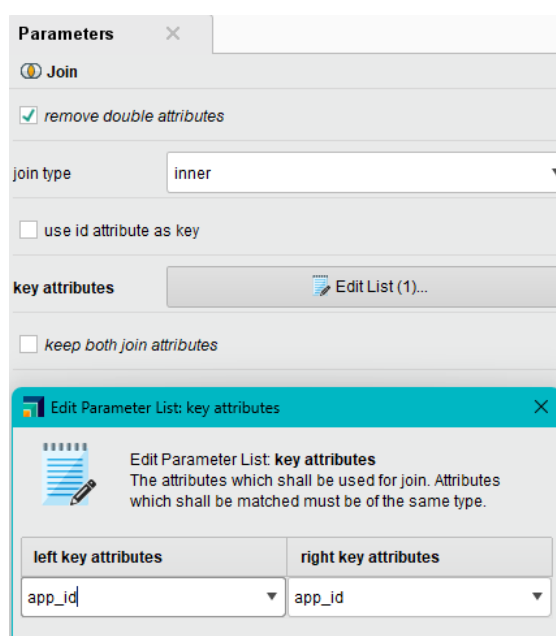
Figura 6 - Design da junção da tabela “games” e “recommendations”



Edit Parameter List: set roles
This parameter defines new attribute roles.

attribute name	target role
app_id	id

Figura 7 - Lista de parâmetros dos papéis (“games” e “recommendations”)



Parameters

Join

☒ remove double attributes

join type: inner

☐ use id attribute as key

key attributes: Edit List (1)...

☐ keep both join attributes

Edit Parameter List: key attributes
The attributes which shall be used for join. Attributes which shall be matched must be of the same type.

left key attributes	right key attributes
app_id	app_id

Figura 8 - Lista de parâmetros dos atributos chave (“games” e “recommendations”)

Após a junção de duas tabelas criou-se um segundo processo que iria juntar o resultado inicial da junção das duas tabelas (“games” e “recommendations”) com a tabela “users”. Para isso o set role de cada uma iria ser a “user_id”, pois é o ID comum dessas tabelas. Por fim foi então realizado um “join” para se juntar as tabelas sendo que queremos que este tenha em conta que para cada ID do user tem um jogo diferente e uma classificação diferente, posto isto é necessário identificar os atributos chave do lado direito e esquerdo por forma a nenhum ficar perdido.

Process

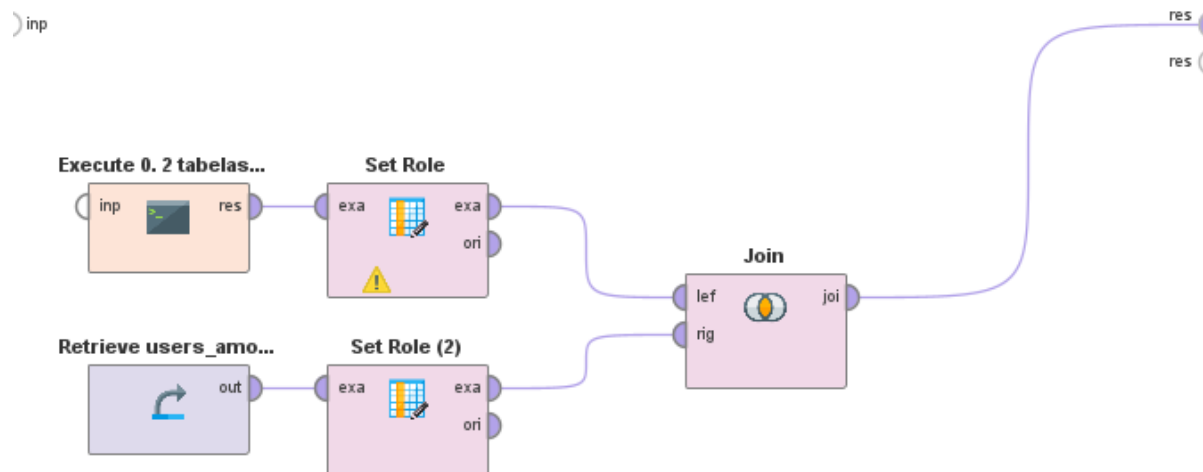


Figura 9 - Design da junção da tabela “games”, “recommendations” e “users”

The screenshot shows a dialog box titled 'Edit Parameter List: set roles'. It contains a text area with the text 'Edit Parameter List: set roles' and 'This parameter defines new attribute roles.' Below this is a table with two columns: 'attribute name' and 'target role'.

attribute name	target role
user_id	id

Figura 10 - Lista de parâmetros dos papéis (“games”, “recommendations” e “users”)

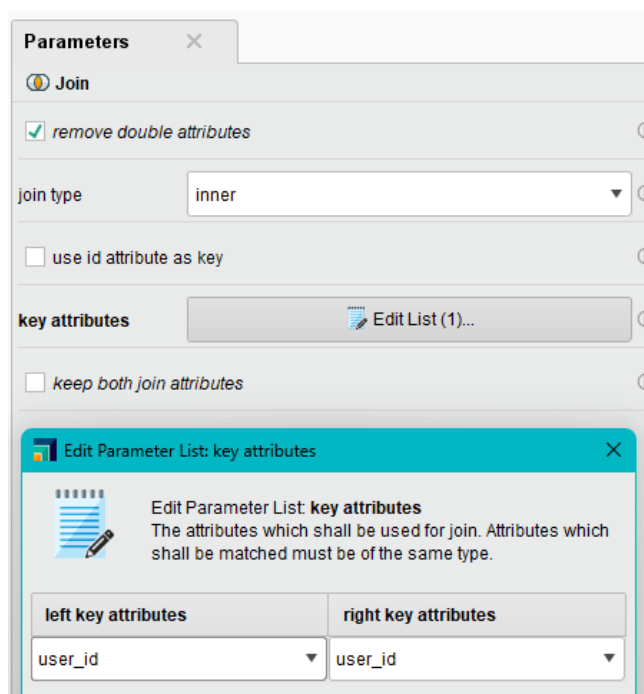


Figura 11 - Lista de parâmetros dos atributos chave ("games", "recommendations" e "users")

Após esses processos decidimos organizar o título dos jogos por ordem alfabética através do operador "sort".

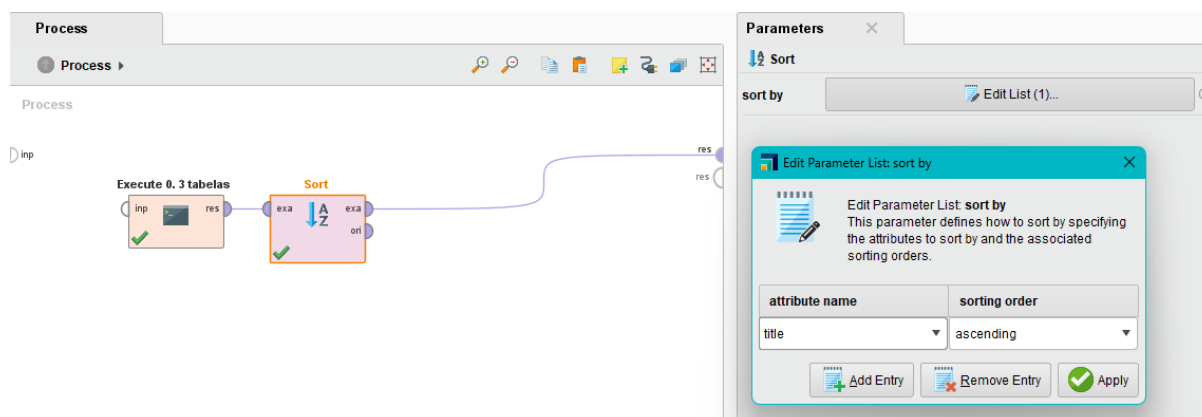


Figura 12 - Design da organização do título

No dataset original alguns atributos estavam em nominal (true or false) e decidimos converter esses mesmos atributos (win, mac, linux, is_recommended) para números inteiros únicos, ou seja, no caso de true e false, true = 1 e false=0. No caso de uma escala qualitativa do rating onde inicialmente antes da junção das tabelas tínhamos 9 classificações, [Figura 13](#), após a junção passamos apenas a ter 4, por isso o rating ficou definido de 0 a 4, [Figura 14](#).

Index	Nominal value	Absolute count	Fraction
1	Very Positive	110	0.604
2	Overwhelmingly Positive	42	0.231
3	Mostly Positive	24	0.132
4	Mixed	6	0.033
5	Mostly Negative	0	0
6	Negative	0	0
7	Overwhelmingly Negative	0	0
8	Positive	0	0
9	Very Negative	0	0

Figura 13 - Rating

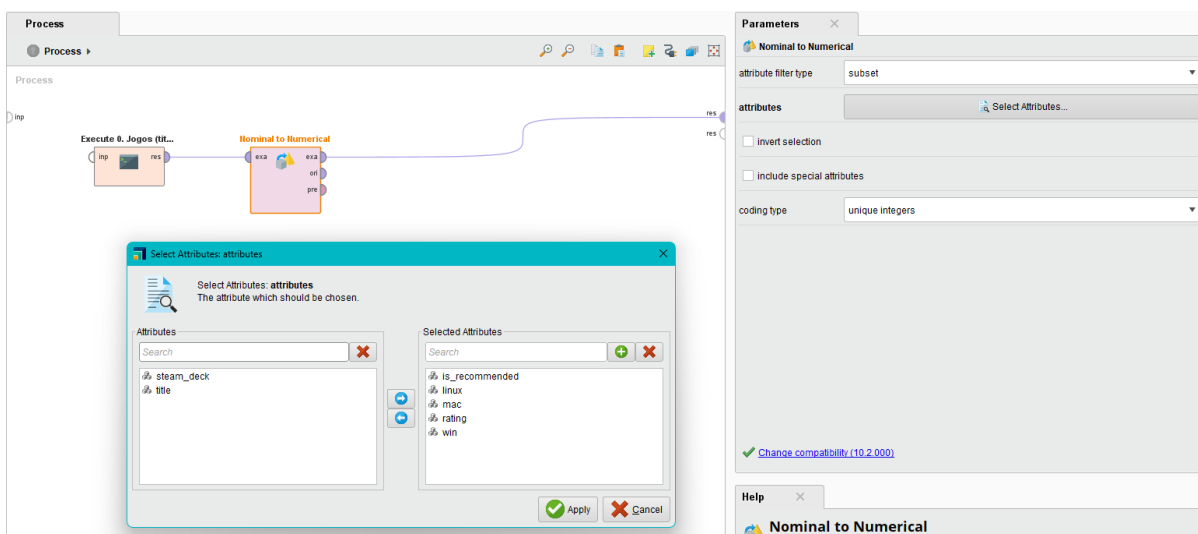


Figura 14 - Conversão de Nominal para Numérico

5.2 Modelação

Depois de se ter atingido a definição dos objetivos de análise e a concretização da análise dos dados, inicia-se a etapa da Modelação, na qual se escolhem e aplicam-se técnicas para a visualização dos dados anteriores, percebendo dessa forma como esses dados se podem correlacionar para retirar certas conclusões.

Desta forma, todos os procedimentos realizados nas fases anteriores evidenciam agora um valor substancial, uma vez que a análise prévia e correta dos dados simplifica o processo de modelação. Neste ponto, já se têm conhecimento dos atributos relevantes, eliminando a presença de dados que poderiam influenciar adversamente o resultado desejado.

5.2.1 Árvore de Decisão

O modelo de visualização escolhido foi o modelo de árvore de decisão, uma vez que este permite uma análise mais aprofundada dos dados em questão. Dado que a nossa questão central está relacionada com a identificação das características dos jogos que são mais previsíveis de serem jogados tendo em contas as classificações feitas pelos utilizadores, acreditamos que este modelo seria eficaz para atingir o objetivo desejado. Através da árvore de decisão é possível perceber as diferentes opções existentes mediante o grau de satisfação em cada variável (jogo), indicando as condições que levam aos diferentes valores da nossa variável de interesse.

Para criar o modelo desejado, foi utilizada a ferramenta RapidMiner. Para tal, foi necessário carregar os dados na plataforma antes de dar continuidade às etapas subsequentes do trabalho.

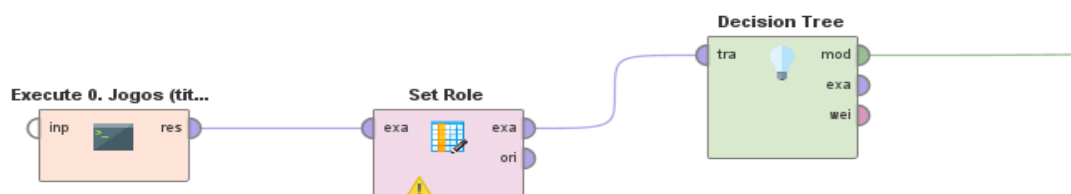


Figura 15 - Criação do Modelo "Decision Tree"

Posto isto, foi necessário inserir um "Set Role" para que fosse possível alterar a função de um ou mais atributos. Neste caso, os atributos "positive_ratio" e "title", Tabela 1, foram definidos como label, ou seja, as nossas funções alvo para encontrar as classificações dos utilizadores e os títulos dos jogos mediante a classificação.

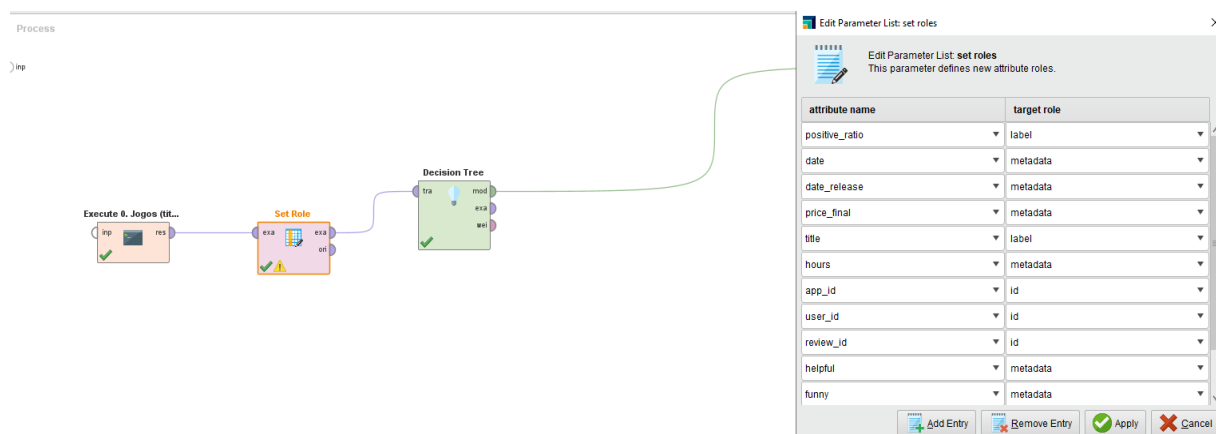


Figura 16 - Atribuição do Label aos atributos "positive_ratio" e "title"

Após a conclusão desse processo, chega-se à utilização do operador "Decision Tree", que gera uma árvore de decisão aplicável à classificação e regressão dos dados.

A determinação dos parâmetros nesta etapa do processo representou uma atividade sujeita a várias iterações e ajustes, visando compreender quais parâmetros melhor se adequavam ao modelo em questão. As modificações realizadas nos parâmetros pré-definidos incluíram principalmente a "Maximal Depth", cujo valor foi ajustado para 3, e a "Confidence", cujo valor foi alterado de 0.1 para 0.5.

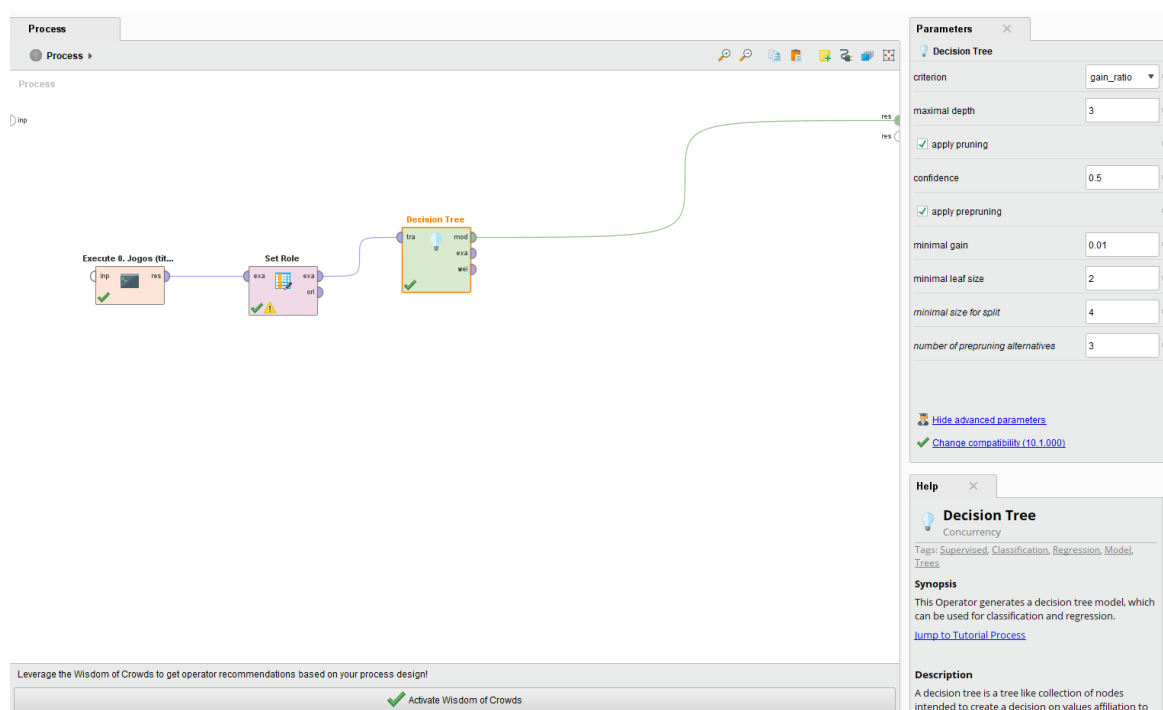


Figura 17 - Definição dos Parâmetros do Modelo

Todas as modificações realizadas visaram alcançar o conteúdo desejado na árvore de decisão final e facilitar a compreensão da mesma. Uma análise mais detalhada de todo esse processo será realizado na fase de avaliação de resultados.

Ao executar o procedimento mencionado anteriormente, a árvore de decisão fornecida pelo RapidMiner foi a seguinte:

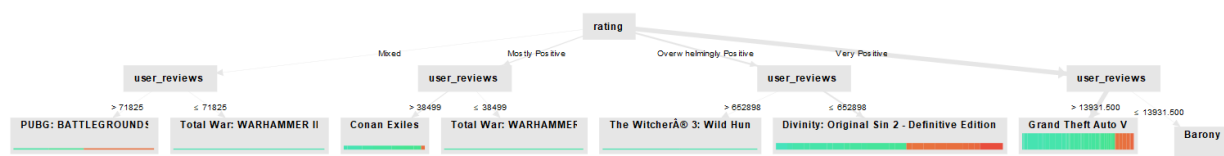


Figura 18 - Resultado da Árvore de Decisão

5.3 Avaliação

As árvores de decisão têm como principal propósito apresentar as possíveis escolhas para resolver um determinado problema. A principal vantagem desse método reside na sua facilidade de compreensão e interpretação. Essas características foram os motivos que orientaram a nossa escolha por este método. Considerando que o nosso problema consistia principalmente em identificar os fatores mais influentes no tipo de risco, a árvore de decisão revelou-se apropriada, pois para cada característica apresentava as condições que levavam a um bom ou mau resultado.

Para criar uma árvore de decisão conforme desejado, foi necessário experimentar diferentes parâmetros, como mencionado anteriormente. O objetivo principal era evitar que a árvore se tornasse muito detalhada (com muitos ramos) ou muito genérica. A intenção era representar de maneira interpretável cada um dos atributos.

Tal como foi dito anteriormente, o parâmetro "Maximal Depth", cujo valor foi ajustado para 3, e a "Confidence", cujo valor foi alterado de 0.1 para 0.5. Posto isto, todos os outros parâmetros foram mantidos com os valores padrões.

Explicado todo o processo de escolha e desenvolvimento do nosso modelo, segue-se a avaliação dos resultados do nosso modelo. A escala de avaliação utilizada foi limitada as categorias "Very Positive", "Overwhelmingly Positive", "Mostly Positive" e "Mixed", através desta foi possível obter insights valiosos sobre como o jogo iria, expectavelmente, ser recebido pela comunidade. Cada escala é dividida em dois, onde na primeira linha apresentamos os números de classificações superiores a um certo número, e na segunda linha o número de avaliações (*reviews*) de utilizadores inferiores ou iguais a esse número. Vamos explorar o significado de cada uma dessas categorias, de forma hierárquica, do maior para o menor valor de satisfação:

- Very Positive (Extremamente Positivo): Esta classificação indica que a grande maioria dos utilizadores avaliou o jogo de forma extremamente positiva. Os jogadores expressaram alto grau de satisfação, destacando elementos como a "jogabilidade", gráficos, história e outros aspetos positivos.
- Overwhelmingly Positive (Muito Positivo): Uma classificação ligeiramente inferior a "Very Positive", contudo, geralmente positiva. Isso sugere que a esmagadora maioria dos jogadores teve uma experiência extremamente positiva com o jogo, indicando uma recepção excecionalmente favorável.
- Mostly Positive (Maioritariamente Positivo): Esta classificação implica que a maioria dos utilizadores avaliou o jogo de maneira positiva, embora possam existir algumas críticas ou ressalvas. Em geral, a resposta é positiva, mas há uma presença de opiniões variadas.
- Mixed (Misto): Quando um jogo recebe a classificação "Mixed", isso sugere uma variedade de opiniões entre os utilizadores bastante acentuada. Pode haver avaliações positivas e negativas de forma equilibrada, indicando uma recepção mais ambígua em relação ao jogo.

A tabela 2 será elaborada para apresentar a informação anterior de uma maneira que facilite a interpretação visual:

Tipo de Rating	Situação
Very positive	Número de reviews de utilizadores superiores a 13931.500, quanto ao jogo "Grand Theft Auto V".
	Número de reviews de utilizadores inferiores ou iguais a 13931.500, quanto ao jogo Barony.
Overwhelming positive	Número de reviews de utilizadores superiores a 652898, quanto ao jogo "The Witcher 3: Wild Hunt"
	Número de reviews de utilizadores inferiores ou iguais a 652898, quanto ao jogo "Divinity: Original Sin 2 – Definitive Edition"
Mostly positive	Número de reviews de utilizadores superiores a 38499, quanto ao jogo "Conan Exiles"
	Número de reviews de utilizadores inferiores ou iguais a 38499, quanto ao jogo "Total War: Warhammer"
Mixed	Número de reviews de utilizadores superiores a 71825, quanto ao jogo "PUBG: BATTLEGROUNDS"
	Número de reviews de utilizadores inferiores ou iguais a 71825, quanto ao jogo "Total War: Warhammer II"

Tabela 2 - Interpretação da situação conforme o Tipo de Rating

No contexto do nosso caso, identificamos que o jogo mais provável de ser escolhido por um jogador é o GTA V, devido a ter alcançado a classificação mais elevada, categorizada como "Very Positive". Este título manteve sua popularidade mesmo diante de uma considerável quantidade de análises por parte dos utilizadores, totalizando 13931.500 avaliações. Pelo que o "Barony" não teve a mesma quantidade de votos, mas conseguiu, mesmo assim, manter-se com uma classificação muito boa.

Relativamente à classificação "Overwhelming positive", o jogo "The Witcher 3: Wild Hunt" mantave-se nesta classificação ao agregar 652898 reviews, já o "Divinity: Original Sin 2 – Definitive Edition" obteve menos que 652898 reviews.

Quanto à classificação "Mostly positive", o jogo "Conan Exiles" teve mais de 38499 reviews, e o jogo "Total War: Warhammer" teve um número inferior a 38499 reviews.

Por fim, o "Mixed" apresenta o jogo "PUBG: BATTLEGROUNDS" onde consegue ter mais de 71825 reviews, enquanto o "Total War: Warhammer II" tem menos de desse número.

O problema que apresentado neste trabalho está relacionado com a influência do grau de satisfação expresso em análises de jogos na probabilidade de outros utilizadores escolherem jogar o mesmo título. Noutras palavras, a satisfação manifestada nas avaliações de um jogo, categorizadas em escalas como "Mixed", "Overwhelmingly Positive", e "Very Positive", tem um impacto significativo na decisão de outros jogadores experimentarem ou não esse jogo em particular.

Quando as análises de um jogo indicam uma classificação elevada, como "Very Positive" ou "Overwhelmingly Positive", isso sugere que a maioria dos jogadores teve uma experiência altamente

satisfatória. Essa avaliação positiva pode influenciar positivamente a decisão de outros jogadores, aumentando a probabilidade de escolherem jogar o mesmo jogo.

Por outro lado, se as análises variam entre "Mixed" e "Mostly Positive", isso indica uma recepção geralmente favorável, mas com algumas críticas ou opiniões mistas. Nesse caso, a probabilidade de outros utilizadores escolherem jogar o jogo pode ser influenciada por aspetos menos positivos mencionados nas análises.

Essa dinâmica destaca a importância das avaliações de jogos como um fator influenciador na decisão dos utilizadores adquirirem e jogarem um determinado jogo. A confiança na satisfação expressa pelos jogadores pode desempenhar um papel crucial na escolha de um jogo, tornando as análises uma ferramenta valiosa para a comunidade de jogadores ao explorar novos títulos. Tornando-se agora possível, devido ao apoio da árvore de decisão, encontrar um padrão de resposta ao problema.

6 Conclusão

Concluimos assim com este projeto que foi possível identificar e cumprir o objetivo proposto que foi: através das avaliações dos utilizadores verificar os jogos que continham melhores avaliações.

As principais dificuldades sentidas ao longo deste trabalho foi essencialmente a construção de um problema que se adequasse às características do dataset, tendo sido este ultrapassado. Posto isto conseguimos ganhar mais proficiência na ferramenta RapidMiner.

Em suma, este trabalho prático demonstrou ser benéfico ao proporcionar uma aplicação mais detalhada dos conceitos abordados na cadeira de Análise e Visualização de Dados, tendo aprofundado e consolidado melhor tanto as aprendizagens práticas como as teóricas.

Referências

- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2020). *CRISP-DM 1.0*. Obtido de Step-by-step data mining guide: https://s2.smu.edu/tfomby/eco5385_eco6380/data/SPSS/CRISPWP-0800%20Data%20Mining%20Standards.pdf
- How to Join Your Data [Tips + Tricks]*. (08 de 09 de 2016). Obtido de RapidMiner: <https://rapidminer.com/blog/tips-tricks-different-ways-to-join-data/>
- Moreira, J. M. (2023). *Data Mining*. Obtido de Moodle: https://moodle2324.up.pt/pluginfile.php/105276/mod_resource/content/2/1%20-%20CRISP%20DM.pptx.pdf
- Replace Missing Values*. (2023). Obtido de RapidMiner: https://docs.rapidminer.com/latest/studio/operators/cleansing/missing/replace_missing_values.html

Anexos

Anexo 1













Name	Type	Missing	Statistics	Filter (22 / 22 attributes)
app_id	Integer	0	 Min: 570, Max: 1967080, Average: 559978.681, Deviation: 464045.050	
user_id	Integer	0	 Min: 27862, Max: 13514834, Average: 7620401.330, Deviation: 3477239.547	
win	Integer	0	 Min: 1, Max: 1, Average: 1, Deviation: 0	
mac	Real	0	 Min: 0, Max: 1, Average: 0.363, Deviation: 0.482	
linux	Real	0	 Min: 0, Max: 1, Average: 0.236, Deviation: 0.426	
rating	Real	0	 Min: 0, Max: 4, Average: 1.385, Deviation: 1.764	
is_recommended	Real	0	 Min: 0, Max: 1, Average: 0.143, Deviation: 0.356	
title	Nominal	0	 Least: 8Y2RabbitWY2 (0), Most: Grand Theft Auto V (10), Grand Theft Auto V (10), Conan Exiles (5), Counter-Strike: Global Offensive (5), ... [48178 more] Details...	
date_release	Date-time	0	 Earliest date: Oct 19, 2010 12:00 AM, Latest date: Mar 13, 2023 12:00 AM, Duration: 4528d 1h 0m 0s	
positive_ratio	Integer	0	 Min: 55, Max: 98, Average: 87.121, Deviation: 8.487	
user_reviews	Integer	0	 Min: 1838, Max: 7494460, Average: 533510.786, Deviation: 1258563.581	
price_final	Real	0	 Min: 0, Max: 70, Average: 23.151, Deviation: 18.216	
price_original	Real	0	 Min: 0, Max: 0, Average: 0, Deviation: 0	
discount	Real	0	 Min: 0, Max: 0, Average: 0, Deviation: 0	
steam_deck	Nominal	0	 Least: false (0), Most: true (152), true (152), false (0), Details...	
helpful	Integer	0	 Min: 0, Max: 50, Average: 1.648, Deviation: 5.503	
funny	Integer	0	 Min: 0, Max: 55, Average: 0.643, Deviation: 4.373	
date	Date-time	0	 Earliest date: Sep 8, 2012 12:00 AM, Latest date: Dec 30, 2022 12:00 AM, Duration: 3765d 1h 0m 0s	
hours	Real	0	 Min: 0.400, Max: 968.700, Average: 227.923, Deviation: 251.597	
review_id	Integer	0	 Min: 971, Max: 50714, Average: 25532.802, Deviation: 14875.270	
products	Integer	0	 Min: 0, Max: 5166, Average: 246.418, Deviation: 471.725	
reviews	Integer	0	 Min: 1, Max: 82, Average: 8.462, Deviation: 13.655	

Figura 19 - Estatísticas relativas aos Atributos