Universidade do Minho
Escola de Engenharia

METI - Emulação e Simulação de Redes de Telecomunicações

Trabalho individual de Inteligência Artificial

Aluno:

Bruno Miguel Fernandes Araújo - pg55806

Docente:

Dalila Alves Durães

23 de dezembro de 2024

Conteúdo

Lista de Figuras

1 Introdução

Foi proposto o estudo de um dataset relativo á venda de video-jogos no mundo, com este tinhamos de , usando a ferramenta KNIME, responder a várias perguntas e usar pelo menos três algoritmos de inteligência artificial com a intenção destes preverem resultados e podermos compara-los ao nível da sua precisão, concluindo qual deles é o melhor.

2 Análise dos Dados

Foram dados dois datasets relativamente às vendas de video-jogos no mundo, um contém o nome e a descrição de cada coluna do dataset que será estudado.

A estrutura do dataset a ser abordado tem as seguintes 14 colunas:

1. **img** - URL para a arte da caixa em vgchartz.com.
2. **title**- Titulo do jogo.
3. **console**- Consola para qual o jogo foi lançado.
4. **genre**- Genero do jogo.
5. **publisher**- Nome de quem publicou o jogo.
6. **developer**- Nome de quem desenvolveu o jogo.
7. **critics_score**- Avaliação do jogo no Metacritic (de 0 a 10).
8. **total_sales**- Numero de vendas mundialmente (em milhões)
9. **na_sales**- Numero de vendas no norte da América (em milhões)
10. **jp_sales**- Numero de vendas no Japão (em milhões)
11. **pal_sales**- Numero de vendas na Europa e na Africa (em milhões)
12. **other_sales**- Vendas no resto do mundo (em milhões)
13. **release_date**- Data do lançamento do jogo
14. **last_update**- Ultima data em que a informação foi atualizada.

Começei por ler a informação que se encontra no dataset, usando o nodo **Csv Reader**. De seguida para analisar a existência de missing values usei o nodo de **Statistics View** e foi possível observar que realmente existem imensos missing values.

Foram também usados outros nodos para a análise de outras componentes:

1. **Linear Correlation**: Saber as correlações entre as colunas (útil para os AI's).
2. **Box Plot**: Verificar se existem outliers (deu para concluir que existem alguns).
3. **Data Explorer** e **Scatter Plot**: Outras formas de visualizar e analisar a informação.

3 Exploração dos Dados

Juntamente com o dataset foram propostas algumas perguntas que seriam interessantes serem respondidas usando o **KNIME**. Analisarei individualmente a metodologia usada para desenvolver uma resposta para estas e para uma extra que achei pertinente.

3.1 A industria tem crescido com o tempo?

Nesta questão as únicas colunas relevantes são a das vendas mundialmente (total_sales) e a das datas do lançamento do jogo (release_date) (não existe necessariamente uma data que reflita as vendas num ano então acho que esta componente seria a mais próxima).

Então comecei por usar o nodo **Column filter** e excluir todas as colunas exceto essas duas.

Agora ,é importante remover os missing values, dando uso ao nodo **Missing Value** optei por remover as linhas que têm falta do valor, ficando então com 18832 linhas.

Ordenei pela data e alterei a string (**String Manipulation**) dando uso da função **regexreplace** com a expressão regular ”([0-9]+)([0-9-]+)”onde mantenho apenas o primeiro grupo da expressão (correspondente ao ano).

Por fim agrupasse por ano (**GroupBy**) somando as total_sales e representei num grafico de barras(**Bar Chart**).

Obtemos o seguinte resultado:

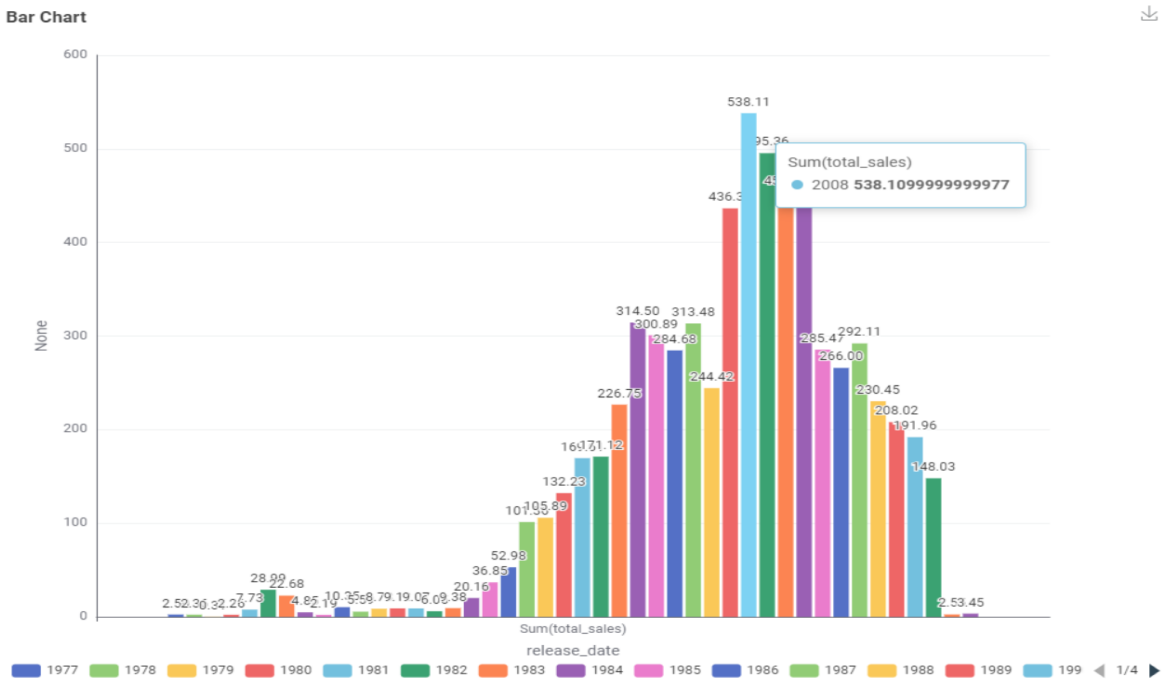

Figura 1: Representação da evolução das vendas por ano.

Infelizmente, não dá para ter uma representação perfeita pois não existe uma variável que represente o número de vendas num ano.

Mas com a informação dada, seria seguro concluir que houve uma subida e uma descida nas vendas ao longo do tempo, e que os jogos lançados em **2008** foram os que venderam mais até agora.

3.2 As consolas especializam-se em algum género de jogos?

Para esta questão, a resolução no **KNIME** é muito simples, filtrei as colunas (**Column Filter**) mantendo apenas as relevantes para este problema, a **console** e a **genre**, e por fim simplesmente agrupei pelo nome da consola contando o número de genres únicos.

Representei o resultado novamente num gráfico de barras (**Bar Chart**) , obtendo o seguinte:

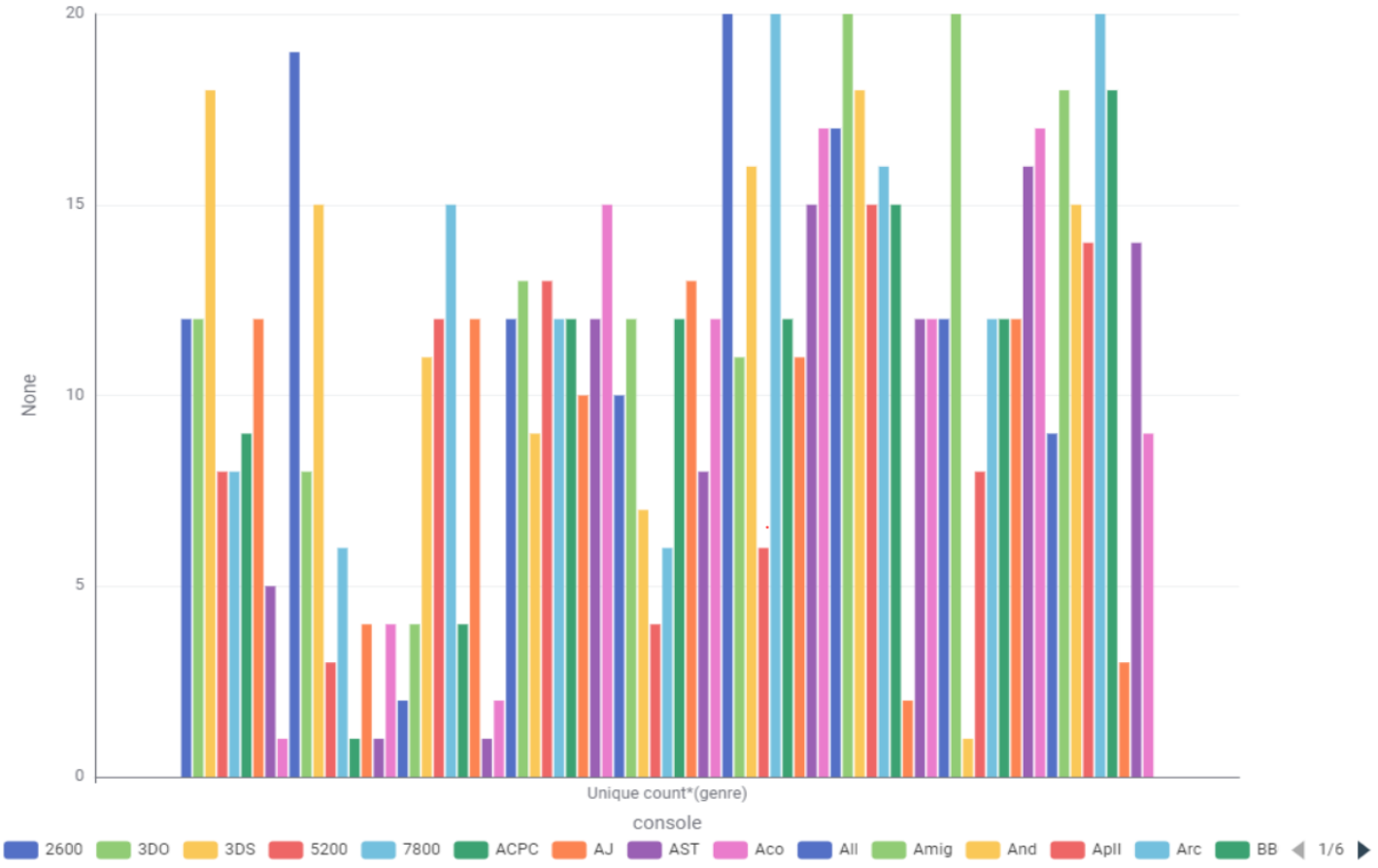

Figura 2: Representação das consolas com a quantidade de generos de jogos diferentes.

Podemos ver algumas consolas com jogos que são apenas de um género, o que significa que estas especializam-se neste (Por exemplo a **ACO**), assim como temos outras com 20 géneros , que é o máximo número de géneros possíveis neste dataset, logo estas aceitam jogos de qualquer género.

3.3 Que jogos é que deram ”flop”numa região e noutras não?

Nesta questão , é necessário observar apenas as colunas com o nome do jogo e tanto o número de vendas por região como na totalidade.

Fiz a devida filtração (**Column Filter**) e eliminei as linhas com missing values (**Missing Value**) diminuindo o número de linhas para 2222 depois agrupei por nomes, somando o número das vendas, isto pois existem varias ocorrências dos mesmos títulos com consolas diferentes.

Para finalizar, apenas ordenei pelo número de vendas na totalidade de forma a aparecer o mais vendido em primeiro e representei o resultado numa tabela (**Table View**),obtendo o seguinte:

Rows: 1705 | Columns: 6

RowID	title *String*	Sum(total_sales) *Number (double)*	Sum(na_sales) *Number (double)*	Sum(jp_sales) *Number (double)*	Sum(pal_sales) *Number (double)*	Sum(other_sales) *Number (double)*
Row...	Grand Theft Auto V	64.29	26.19	1.66	28.14	8.32
Row...	Call of Duty: Modern ...	28.17	14.61	0.62	10.07	2.87
Row...	Call of Duty: Black Op...	27.66	13.26	0.72	10.2	3.48
Row...	Call of Duty: Black Ops	27.41	15.77	0.59	8.13	2.92
Row...	Call of Duty: Ghosts	25.06	12.88	0.49	8.38	3.32
Row...	Call of Duty: Black Op...	24.41	11.46	0.5	9.01	3.42
Row...	Call of Duty: Modern ...	24.14	13.53	0.46	7.29	2.87
Row...	Grand Theft Auto IV	21.66	11.59	0.58	6.84	2.66
Row...	Call of Duty: Advance...	21.36	10.49	0.35	7.77	2.78
Row...	Call of Duty 4: Moder...	17.18	10.06	0.42	4.68	2.03
Row...	The Elder Scrolls V: S...	16.52	8.34	0.41	5.83	1.94
Row...	Guitar Hero III: Legen...	16.36	11.12	0.04	2.57	2.64
Row...	Grand Theft Auto: Vic...	16.15	8.41	0.47	5.49	1.78
Row...	FIFA 14	14.69	2.43	0.21	10.17	1.89

Figura 3: Representação dos possíveis flops em certas regiões

Podemos observar que realmente certos jogos têm mais sucesso numas regiões do que noutras, por exemplo o **FIFA 14** vendeu muito mais na europa e na africa (região **Pal**), do que no norte da América (região **NA**). Também é possível concluir que o jogo mais vendido de sempre (neste dataset) é o **Grand Theft Auto V**.

3.4 Extra: As avaliações dos críticos afetam as vendas?

Sendo que este dataset tem presente uma coluna de avaliação dos jogos e que existe uma pequena correlação entre este score com o número de vendas na totalidade, achei pertinente colocar esta questão.

Para então responder a esta, segui a mesma lógica que as perguntas anteriores, filtrei as colunas desnecessárias (**Column Filter**) , mantendo apenas o nome do jogo, o rating e o total das vendas.

Obviamente foi necessário corrigir a presença dos missing values e tomei a mesma decisão que nos outros casos, removi as linhas com estes (**Missing Value**) ficando com cerca de 4126 destas.

Agrupei pelo nome (**GroupBy**) onde somei o total das vendas e guardei o máximo dos scores de cada jogo.

Finalizei a resolução, ordenando de forma descendente o número de vendas na totalidade (**Sorter**) e representei o resultado numa tabela (**Table View**).

Rows: 2876 | Columns: 3

RowID	title *String*	Sum(total_sales) *Number (double)*	Max*(critic_score) *Number (double)*
Row...	Call of Duty: Black Ops II	28.08	8.6
Row...	Call of Duty: Modern Warfare 2	25.02	9.5
Row...	Grand Theft Auto IV	22.53	10
Row...	Call of Duty: Advanced Warfare	21.78	9.1
Row...	The Elder Scrolls V: Skyrim	20.51	9.3
Row...	Call of Duty 4: Modern Warfare	18.33	9.6
Row...	Battlefield 3	17.32	8.9
Row...	Guitar Hero III: Legends of Rock	16.36	8.7
Row...	FIFA 15	16.28	8.1
Row...	Grand Theft Auto: Vice City	16.19	9.6
Row...	FIFA 18	16.12	8.3
Row...	Call of Duty: World at War	15.94	8.6
Row...	FIFA 16	15.82	9
Row...	Call of Duty: Black Ops 3	15.09	8.1

Figura 4: Representação da possibilidade da influência do score com as vendas.

Posso seguramente concluir que o score não influencia as vendas, existem jogos ,alguns até da mesma franquia, que têm um número de vendas superior a uma futura entry no franchise apesar de ter um score pior. Por exemplo, as duas primeiras linhas da tabela da figura anterior, dois jogos da franquia **Call of Duty** em que um tem maior número de vendas e pior score que o outro.

4 Tratamento dos Dados para os AI's

Para os dados serem previstos, seriam bom o AI ser treinado com informação que esteja correlacionada, então , após observar o **Linear Correlation**, concluí que seria interessante ,no contexto do trabalho, prever a avaliação através das vendas totais,pelo genero e pela consola.

Segui o mesmo processo das outras perguntas, mantive as colunas importantes (**Column Filter**), removi os missing values (**Missing Value**) removendo as linhas que continham estes.

Para ser usado o Partitioning para os algoritmos preverem o critic score, era necessário converter para string os valores, assim o fiz com o uso do nodo **Double to Integer** seguido de dois **rule engines** (um para as avaliações e outra para o número de vendas) que decidi que seriam uteis para criar ranges facilitando o número de diferentes decisões que os algoritmos necessitam de fazer.

Por fim apenas excluí as colunas antigas e de forma que estas ficam substituídas com os ranges (**Column Filter**).

5 Algoritmos de Inteligência Artificial

Após então o tratamento do dados, podemos passar para treino e aplicação dos algoritmos, usei 3 algoritmos que se encontram ligados a um **Partitioning** que irá colocar 70% da informação para treino.

5.1 Decision Tree

O decision tree que foi usado nas aulas e que usa os nodos **Decision Tree Learner** e o **Decision Tree Predictor**. Com um scorer que compara os a coluna prevista e a inicial das avaliações , obtive **54.63**%.

5.2 XBOOST Tree

O XBOOST Tree ,recomendado pela professora e que usa os nodos **XBOOST Tree Ensemble Learner** e o **XBOOST Predictor**, onde coloquei para este aprender com as colunas previamente ditas, genero, consola e Sales-Range. Usei um scorer que compara os a coluna prevista e a inicial das avaliações , obtive **53.31**%.

5.3 Random Forest

O Random Forest que foi usandos nas aulas e que usa os nodos **Random Forest Learnerd** e o **Random Forest Predictor** onde coloquei para este aprender com as colunas previamente ditas, genero, consola e Sales-Range. Usei um scorer que compara os a coluna prevista e a inicial das avaliações , obtive **56.03**%.

5.4 Comparação dos Resultados

Comparando então os resultados, podemos então concluir que o **Random Forest** foi o algoritmo que preveu melhor, mesmo alterando as tabelas ou outras componentes, este foi sempre consistentemente o melhor dos três.

6 Conclusão

Todos os tópicos deste projeto foram resolvidos, explorei o dataset respondendo a diferentes questões propostas usando os nodos do KNIME que foram utilizados durantes as aulas. Usei três algoritmos de AI , criando uma variedade de resultados interessante e onde tive de usar conceitos que aprendi na minha licenciatura (Expressões regulares, um website útil para o teste destas é o ”regex101”).

A inteligência artificial está a revolucionar o mundo e foi muito interessante tocar nestes diferentes métodos que poderão estar até agora a ser usados em projetos de grande escala, é pena é que não foram dados em código mas sim na ferramenta KNIME, não deteorou a minha motivação , mas realmente seria engraçado saber como estes seriam em código puro.