Lían König e Pedro Schenkel
Inteligência Artificial e Computacional
20 de novembro de 2025

Modelagem de Classificação Supervisionada de Ratings ESRB

1 Introdução

Para a realização deste trabalho, foi selecionado o dataset “Video Games Rating By 'ESRB'”. O endereço para acesso ao conjunto de dados está localizado na seção de referências ao final do relatório [1]. O endereço para o acesso ao repositório com os arquivos utilizados no trabalho também se encontra na seção de referências [2].

1.1 O que é a classificação ESRB

O Entertainment Software Rating Board (ESRB) é a organização que analisa, decide e coloca as classificações etárias indicativas para jogos eletrônicos. Sua meta é ajudas os consumidores em determinar o conteúdo de um jogo e para quem ele é intencionado. A classificação do jogo é exibida na sua caixa em anúncios e nos sites dos jogos [3].

1.2 Propósito

O dataset “Video Games Rating By 'ESRB'” reúne informações sobre diversas características de conteúdo presentes em videogames, como nudez, uso de drogas, violência, entre outros. Além disso, inclui a classificação etária atribuída pela ESRB. O objetivo principal é prever essa classificação etária com base nas características de conteúdo, sendo a coluna de classificação (ESRB) a variável alvo do conjunto de dados.

2 Levantamento e caracterização

2.1 Quantidade de atributos

O dataset contém 34 atributos, dos quais os 2 primeiros (title e console) foram desconsiderados na etapa de modelagem de classificação supervisionada. O último atributo corresponde à variável alvo.

2.2 Atributos e tipagem

Os 32 atributos utilizados na modelagem são: alcohol_reference, animated_blood, blood, blood_and_gore, cartoon_violence, crude_humor, drug_reference, fantasy_violence, intense_violence, language, lyrics, mature_humor, mild_blood, mild_cartoon_violence, mild_fantasy_violence, mild_language, mild_lyrics, mild_suggestive_themes,

mild_violence, no_descriptors, nudity, partial_nudity, sexual_content, sexual_themes, simulated_gambling, strong_language, strong_sexual_content, suggestive_themes, use_of_alcohol, use_of_drugs_and_alcohol, violence e esrb_rating.

Todos esses atributos são do tipo inteiro e assumem apenas os valores 0 ou 1, indicando respectivamente ausência ou presença da característica no jogo. A única exceção é o atributo esrb_rating, que pode assumir os valores E, ET, T ou M, representando as diferentes classificações etárias.

ESRB Rating	Descrição	Faixa etária
E	Everyone	Indicado para todas as idades
ET (E10+)	Everyone 10+	Adequado para pessoas acima de 10 anos
T	Teen	Recomendado para maiores de 13 anos
M	Mature	Recomendado para maiores de 17 anos

2.3 Valores faltando

O dataset não possui nenhum valor faltando.

2.4 Atributo mais desbalanceado

A tabela a seguir mostra os atributos mais desbalanceados do dataset:

Atributo	Não possui	Possui
animated_blood	1876	19
mature_humor	1873	22
nudity	1867	28
partial_nudity	1870	25

2.5 Quantos exemplos o dataset possui

O dataset é composto por dois arquivos: um destinado ao treinamento e outro ao teste do modelo. O arquivo de treinamento contém 1.895 linhas, enquanto o arquivo de teste possui 500 linhas, cada uma representando um exemplo distinto.

3 Pré-processamento

Não foi realizado nenhum tipo de pré-processamento nos dados; todos os valores foram utilizados exatamente como disponibilizados no repositório do Kaggle. Dessa forma, não houve normalização, padronização, balanceamento de classes ou transformação dos atributos. A única adaptação feita antes da modelagem foi a exclusão dos atributos title e console, que não contribuíam diretamente para a classificação

4 Resultados e conclusão

4.1 Tabela com métricas de performance

Algoritmo	Acurácia	Precisão	Recall	F-Score	Tempo de processamento
SVC	85,40%	0.8727	0.8449	0.8557	3,4s
RandomForestClassifier	85,80%	0.8734	0.8512	0.8576	5,1s
LogisticRegression	81,80%	0.8444	0.8077	0.8190	0,6s
DecisionTreeClassifier	83,20%	0.8461	0.8287	0.8332	0,1s

4.2 Tabela com métricas de acurácia para atributo alvo

Algoritmo	Acurácia			
	E	ET	T	M
SVC	95,96%	86,99%	87,50%	78,64%
RandomForestClassifier	94,23%	85,07%	89,55%	80,51%
LogisticRegression	95,96%	80,45%	87,10%	74,27%
DecisionTreeClassifier	94,06%	79,72%	85,51%	79,14%
Média	95,05%	83.05%	87,41%	78,14%

4.3 Análise de performance

Considerando apenas a acurácia dos algoritmos e desconsiderando o tempo de processamento, o RandomForestClassifier apresentou o melhor desempenho, alcançando 85,80% de acurácia. Por outro lado, o LogisticRegression obteve o pior resultado entre os modelos testados, com 81,80% de acurácia.

Ao analisar somente o tempo de processamento dos algoritmos, sem levar em conta a acurácia, o DecisionTreeClassifier foi o mais eficiente, concluindo a execução em 0,1 segundos. Já o RandomForestClassifier apresentou o pior tempo, levando 5,1 segundos para ser processado.

No entanto, ao considerar todas as métricas de desempenho em conjunto, o algoritmo que mais se destacou foi o SVC. Ele atingiu 85,40% de acurácia, apenas 0,40% abaixo da melhor performance, e ainda apresentou um tempo de processamento menor que o do RandomForestClassifier, executando em 3,4 segundos, ou seja, 1,7 segundo mais rápido.

4.4 Análise de acurácia para atributo alvo

Ao analisarmos o desempenho dos modelos por classe da variável target, observamos que o valor de rating que apresentou os melhores resultados médios foi “E”, alcançando uma acurácia de 95,05%. Isso indica que os algoritmos foram bastante eficientes em identificar corretamente jogos classificados para todas as idades.

Por outro lado, o rating com pior desempenho médio foi “M”, com uma acurácia de 78,14%, sugerindo maior dificuldade dos modelos em distinguir corretamente jogos destinados ao público adulto.

A ordem de acurácia média, do melhor para o pior desempenho por classe, ficou da seguinte forma: E (95,05%), T (87,41%), ET (83,05%) e M (78,14%).

5 Referências

[1] https://www.kaggle.com/datasets/imohtn/video-games-rating-by-esrb

[2] https://github.com/zLianK/esrb-ratings-notebook

[3] https://pt.wikipedia.org/wiki/Entertainment_Software_Rating_Board