Supervised Learning: Predicting video games user review scores

Magda Costa – up202207036
Sofia Machado – up202207203

Supervised Learning: Predicting video games user review scores

A análise exploratória de dados e a aplicação de modelos de ‘supervised learning’ de classificação desempenham um papel crucial no campo da ciência de dados, permitindo a extração de insights valiosos e a tomada de decisões baseadas em conjuntos de dados complexos. Neste trabalho, ambicionamos realizar uma análise aprofundada dos dados fornecidos, referentes a videojogos, com o propósito de classificá-los mediante a utilização dos modelos de aprendizado supervisionados.

FORMULAÇÃO DO PROBLEMA

Objetivo

Prever se um videojogo tem boas ou más classificações 'bad', 'mediocre', 'good', 'great'.

DataSet

Base de dados com um total de 5824 inputs

Etapas

Análise de Dados; Pré-processamento de Dados; Análise Exploratória; Classificação; Comparação Resultados

FORMULAÇÃO DO PROBLEMA

DataSet

01 **Id**

02 **Name**

03 **Category**
(e.g,, main game, expansion)

04 **Number of DLCs**

05 **Number of Expansions**

06 **Release Year**

07 **Folows**
(number of people following a game on the IGDB website)

08 **In a Franchise**
(e.g., "Star Wars Racer" → True since it belongs to the Star Wars franchise)

09 **Genre**
(e.g., "Action, Sport")

10 **Platform**
(e.g., "Xbox, PC")

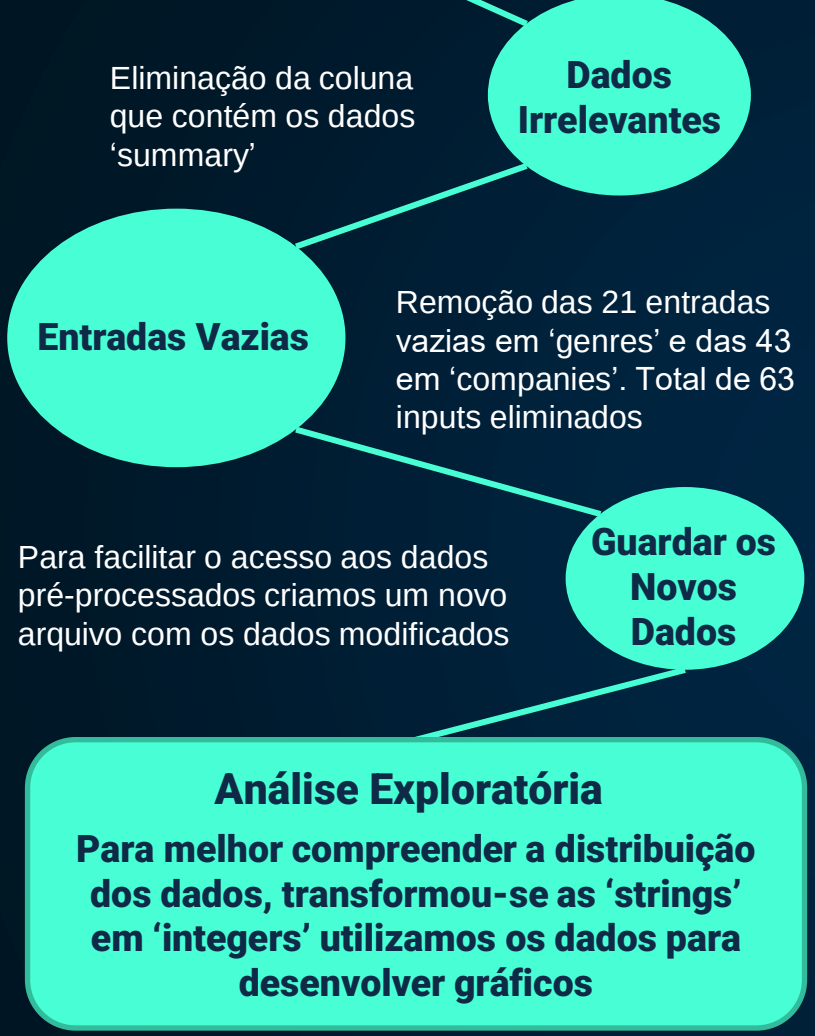
11 **Companies**
(e.g., "Electronic Arts, EA Canada)

12 **Average User Score**
(0 to 100)

13 **Average User Rating**
(bad, mediocre, good or great). Each class represents ~25% of the data.

14 **Number of reviews by users**

15 **Summary**



Pré-Processamento dos Dados

Esta etapa envolve a aplicação de uma série de técnicas e transformações aos dados brutos com o objetivo de prepará-los para a análise e modelagem.

A este processo antecedeu-se uma ‘Análise de Dados’ para uma melhor identificação de erros ou da relevância dos dados.

CLASSIFICAÇÃO

DEFINIR O TARGER target_c = ‘user_rating’

Configuração	Todas as variáveis (incluindo ‘user_score’)
Divisão	Preparação para treinar os modelos
Decision Tree	~ 0,99
K-NN	~ 0,54
Importância Variáveis	‘user_score’ com importância de 1.0

CLASSIFICAÇÃO

Utilizamos todos os dados exceto o ‘user_score’

A remoção do ‘user_score’ aplica-se nos testes seguintes

O ‘user_score’ influencia altamente os resultados

Removemos as variáveis menos relevantes do teste 2

Neste teste não usamos ‘category’, ‘in_francise’, ‘n_dlcs’, ‘n_expansions’ Relevância inferior a 0,08

A remoção dos dados menos relevantes não apresentou resultados muito diferentes

Testamos a relevância do ‘id’ e do ‘nome’ removendo-os

Não utilizamos: ‘category’, ‘in_francise’, ‘n_dlcs’, ‘n_expansions, ‘id’, ‘nome’

A remoção do ‘id’ e do ‘nome’ não apresentou resultados muito diferentes

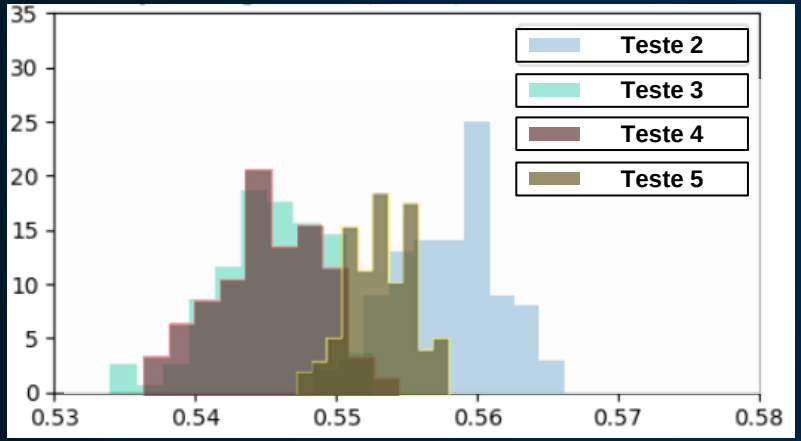
Apenas dados do tipo ‘numerical’

Usamos apenas: ‘id’, ‘n_dlcs’, ‘n_expansions’, ‘year’, ‘follows’, ‘n_user_reviews’

A passagem de ‘string’ para ‘numérico’ não altera os dados de um modo relevante

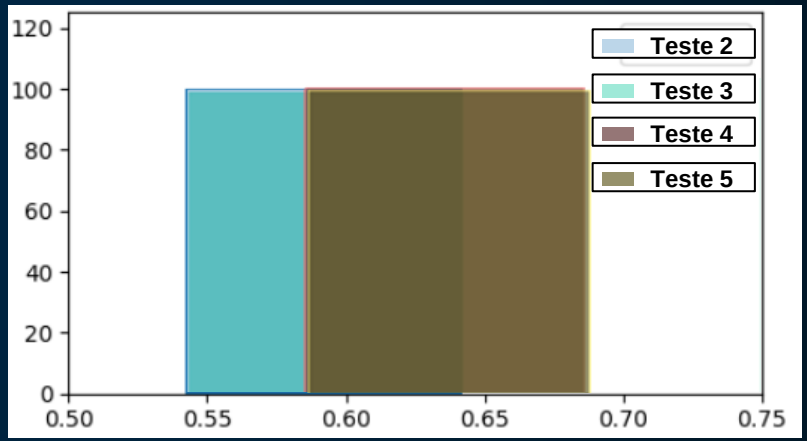
COMPARAÇÃO DOS RESULTADOS

Decision Tree



O Teste 2 foi o que apresentou os melhores resultados. Posteriormente realizando um 'Parameter Tuning' foi possível alcançar uma accuracy de 0,642

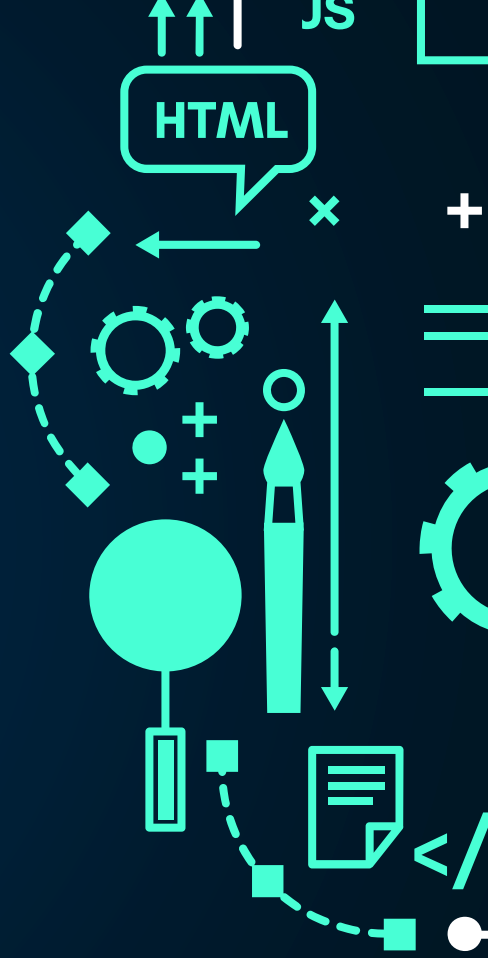
K-NN



Os melhores resultados foram apresentados pelo Teste 5. Posteriormente realizando um 'Parameter Tuning' foi possível alcançar uma accuracy de 0,689

CONCLUSÃO

- Na análise de dados concluímos que nem todos os dados fornecidos são estritamente necessários para a realização do trabalho e que existiam valores com entradas nulas. Fazendo posteriormente a passagem de string para integer.
- Na classificação, realizamos 5 testes, onde tentamos responder a diversas perguntas, como a importância dos dados e o impacto de dados do tipo string. Tendo sido feito o cross-validation para todos os testes exceto o primeiro, de modo a tornar o parameter tuning mais fácil de calcular. Neste foi apenas escolhido o teste com maior precisão para cada caso.
- Ao comparar os testes constatamos que estes não apresentam uma discrepância de valores, sendo todos muito próximos. De um modo geral concluímos que neste caso o melhor método foi o K-NN, já que atingiu valores mais elevados.



Magda Costa – up202207036
Sofia Machado – up202207203