



Universidade Federal de São João del Rei
Campus Tancredo Neves - Ciência da Computação

TRABALHO PRÁTICO 1

Geraldo Arthur Detomi
Oscar Alves Jonson Neto
Pedro Henrique Siqueira De Paula Peixoto
Leonardo da Silva Vieira

São João del-Rei
2025

Sumário

Sumário	2
1. Definição do problema	3
2. Coleta e pré-processamento dos dados	3
2.1 Base de Dados	5
2.2 Pré-processamento dos Dados	5
3. Escolha da técnica	6
4. Desenvolvimento do modelo	13
5. Análise dos resultados	17
6. Conclusão	18

1. Definição do problema

O objetivo deste projeto é desenvolver e avaliar modelos de aprendizado de máquina capazes de prever a avaliação de usuários (*User Rating*) de um jogo digital com base em um conjunto de suas características.

O estudo concentra-se na previsão da nota atribuída a cada jogo, considerando atributos como preço, gênero, plataforma e aspectos técnicos incluindo gráficos, trilha sonora e enredo.

Essa abordagem é especialmente relevante para desenvolvedores e publicadoras, pois permite identificar quais características exercem maior influência na recepção do público, fornecendo informações estratégicas para decisões de design, marketing e posicionamento de produto.

2. Coleta e pré-processamento dos dados

2.1 Base de Dados

Os dados utilizados foram carregados a partir de um arquivo local (*dataset/games.csv*). O conjunto contém informações sobre diversos jogos digitais, abrangendo atributos categóricos (como *Platform* e *Genre*), atributos numéricos (como *Price* e *Game Length (Hours)*) e a variável alvo (*User Rating*).

2.2 Pré-processamento dos Dados

O tratamento dos dados constituiu uma etapa essencial do projeto e foi realizado em quatro etapas principais:

- 1. Leitura e Separação:** O arquivo foi lido utilizando a biblioteca pandas. A coluna alvo (*User Rating*) foi isolada, enquanto colunas consideradas irrelevantes para a tarefa de predição (como *Game Title*, *User Review Text*, *Developer* e *Publisher*) foram removidas.

2. **Transformação Ordinal:** Colunas categóricas que possuem uma ordem intrínseca (como *Graphics Quality* e *Story Quality*) foram mapeadas para valores numéricos (ex: 'Poor' para 0, 'Excellent' para 3) para preservar essa relação hierárquica.
3. **Identificação de Features:** As colunas de atributos foram segregadas em duas listas: *numeric_features* (para dados contínuos ou ordinais) e *categorical_features* (para dados nominais).
4. **Divisão dos Dados:** O dataset foi dividido em conjuntos de treino (80%) e teste (20%) usando a função *train_test_split* do *Scikit-learn*, garantindo uma base de avaliação imparcial para o modelo.

3. Escolha da técnica

Foi adotada a abordagem de **Aprendizado Supervisionado**, especificamente um **problema de Regressão**, pois o objetivo é prever um valor numérico contínuo (a nota do usuário) e o *dataset* possui dados rotulados (jogos com suas respectivas notas).

Para fins de comparação, dois algoritmos de regressão foram selecionados:

- **Regressão Linear:** Um modelo estatístico simples e interpretável, usado como linha de base (*baseline*) para a performance.
- **Random Forest Regressor:** Um modelo de ensemble (baseado em árvores de decisão) robusto, capaz de capturar relações não-lineares complexas nos dados e geralmente apresentando alta precisão.

4. Desenvolvimento do modelo:

Para otimizar o fluxo de trabalho e aplicar as transformações de forma consistente, o desenvolvimento foi encapsulado em Pipelines do Scikit-learn.

- **Transformação de Features:** Foi utilizado um *ColumnTransformer* para aplicar diferentes transformações a diferentes subconjuntos de colunas simultaneamente:

- **Features Numéricas:** (*numeric_transformer*): Foi aplicado o *StandardScaler*, que normaliza as features removendo a média e escalonando para a variância unitária.
- **Features Categóricas:** (*categorical_transformer*): Foi aplicado o *OneHotEncoder* (com *handle_unknown='ignore'*), que converte variáveis categóricas em vetores binários (dummy variables), permitindo que os modelos as processem.
- **Treinamento do Pipeline:** Dois pipelines completos foram criados, cada um contendo o pré-processador (*ColumnTransformer*) e um regressor (*LinearRegression* ou *RandomForestRegressor*).

Ambos os pipelines foram treinados (usando `.fit()`) apenas no conjunto de dados de treino (X_{train} , y_{train}).

5. Análise dos resultados:

A performance de ambos os modelos foi avaliada no conjunto de teste (X_{test} , y_{test}), utilizando três métricas padrão para problemas de regressão:

1. **R-squared (R2):** O coeficiente de determinação. Indica a proporção da variância na nota do usuário que é explicada pelo modelo. Valores mais próximos de 1 indicam um melhor ajuste.
2. **Mean Absolute Error (MAE):** O erro médio absoluto. Indica, em média, "o quanto" (na unidade da nota) as previsões do modelo erraram, sem considerar a direção do erro.
3. **Root Mean Squared Error (RMSE):** A raiz do erro quadrático médio. Similar ao MAE, mas penaliza mais os erros grandes.

Os resultados (R2, MAE e RMSE) foram calculados e exibidos separadamente para a Regressão Linear e para o Random Forest, permitindo uma comparação direta de sua eficácia preditiva.

Além disso, para o modelo Random Forest, foi extraída a importância das features (*feature_importances_*). Esta análise revela quais atributos (ex: *Graphics Quality*, *Price*, *Genre_Action*, etc.) tiveram o maior peso na decisão do modelo ao

prever a nota, fornecendo insights valiosos sobre o que mais impacta a avaliação de um jogo.

Após o treinamento e teste dos modelos, ambos, Regressão Linear e Random Forest Regressor, apresentaram desempenhos bastante satisfatórios, com altos valores de R², indicando forte capacidade de explicação da variância da variável-alvo (User Rating).

Para a Regressão Linear, o coeficiente de determinação R² foi de 0.9769, o que significa que aproximadamente 97,7% da variação nas avaliações dos usuários foi explicada pelo modelo. O erro médio absoluto (MAE) foi de 1.0008, enquanto o erro quadrático médio (RMSE) foi de 1.1575, demonstrando que, em média, as previsões do modelo diferem em cerca de 1 ponto da nota real dos usuários.

Já o modelo de Random Forest apresentou um desempenho muito próximo, com R² = 0.9746, MAE = 1.0326 e RMSE = 1.2120. Embora a regressão linear tenha obtido métricas ligeiramente melhores, a diferença é pequena, o que mostra que ambos os modelos conseguem capturar de forma eficaz as relações entre as características dos jogos e suas avaliações pelos usuários. No entanto, o Random Forest tende a generalizar melhor e ser menos sensível a outliers, o que pode ser vantajoso em cenários com maior variabilidade nos dados.

Além das métricas de desempenho, foi analisada a importância das features no modelo Random Forest. Observa-se que o atributo "Price" é o fator mais relevante para determinar a nota do usuário, com importância de 0.5386, seguido por "Game Length (Hours)" (0.4007) e "Release Year" (0.0232). Esses resultados indicam que o preço e a duração do jogo são determinantes centrais na percepção de valor por parte dos jogadores, pois quanto mais justo o preço e mais extenso o conteúdo, melhor tende a ser a avaliação. Outros fatores como "Story Quality", "Soundtrack Quality" e "Graphics Quality" também aparecem entre as principais variáveis, reforçando a influência dos aspectos técnicos e imersivos na experiência do usuário.

Essa análise evidencia que o modelo foi capaz não apenas de prever com precisão as avaliações, mas também de oferecer insights interpretáveis sobre os fatores que mais impactam a satisfação dos jogadores, fornecendo subsídios relevantes para estratégias de design e marketing de jogos digitais.

```
Dados divididos: 38219 para treino, 9555 para teste.

Resultados (Regressão Linear):
R-squared (R2): 0.9769
Mean Absolute Error (MAE): 1.0008
Root Mean Squared Error (RMSE): 1.1575

Resultados (Random Forest):
R-squared (R2): 0.9746
Mean Absolute Error (MAE): 1.0326
Root Mean Squared Error (RMSE): 1.2120

Top 10 features mais importantes:
| Feature | Importance |
| :----- | :----- |
| num_Price | 0.583663 |
| num_Game Length (Hours) | 0.400608 |
| num_Release Year | 0.00234054 |
| num_Min Number of Players | 0.0019388 |
| num_Story Quality | 0.00112601 |
| num_Soundtrack Quality | 0.00111519 |
| num_Graphics Quality | 0.00110969 |
| cat_Age Group Targeted_All Ages | 0.000396506 |
| cat_Age Group Targeted_Adults | 0.000390816 |
| cat_Platform Mobile | 0.000383964 |
```

Figura 1 – Resultados obtidos pelos modelos de Regressão Linear e Random Forest.

6. Conclusão:

Este trabalho implementou com sucesso um pipeline de machine learning supervisionado para a predição de avaliações de jogos. A utilização de Pipelines e *Column Transformers* garantiu um pré-processamento de dados robusto e replicável.

A avaliação comparativa entre a Regressão Linear e o *Random Forest* permitiu quantificar a diferença de performance entre um modelo base e um modelo de ensemble mais complexo. A análise final da importância das *features*, possibilitada pelo *Random Forest*, validou a capacidade do modelo não apenas de prever, mas também de fornecer *insights* sobre quais fatores são mais relevantes para o sucesso de um jogo digital.