Tópicos Contemporâneos 3
Projeto da Discipina - AV2

Antonio Neto, Davi Cesar, João Ricardo

Ciência da Computação –CESAR School
Avenida, Cais do Apolo, 77, Recife-PE - 50030-22

ANÁLISE DE DADOS DE PARTIDAS RANQUEADAS DE LEAGUE OF LEGENDS

RESUMO

O presente relatório tem como objetivo analisar um conjunto de dados contendo informações completas de partidas ranqueadas de *League of Legends*. A análise envolve limpeza, validação, tratamento de inconsistências, remoção de duplicatas e treinamento de um modelo de aprendizado de máquina Random Forest para prever o vencedor de uma partida com base apenas nos campeões selecionados pelos times. Após a limpeza dos dados, treina-se um classificador com divisão de treino e teste de 70/30. O modelo obteve acurácia de aproximadamente **51,8%**, indicando baixa capacidade preditiva sob as condições utilizadas. O relatório apresenta todas as etapas do processo, bem como conclusões a respeito da qualidade do dataset e da validade do modelo gerado.

1. INTRODUÇÃO

A análise de dados esportivos eletrônicos cresce a cada ano, acompanhando o aumento da popularidade dos *eSports*. O jogo *League of Legends*, um dos títulos competitivos mais consolidados, oferece diversos indicadores de desempenho que podem ser explorados por técnicas de ciência de dados e aprendizado de máquina.

Este trabalho descreve o processo completo de tratamento e análise de um dataset contendo mais de 60 variáveis por partida, visando construir um modelo capaz de prever o vencedor utilizando apenas os campeões selecionados pelos times. O trabalho inclui:

- Inspeção e checagem de nulos e valores inválidos;
- Validação de integridade e consistência;
- Remoção de outliers;
- Eliminação de duplicatas;
- Construção e avaliação de um modelo *Random Forest*.

2. METODOLOGIA

2.1 Importação de bibliotecas e leitura dos dados

Foram utilizadas as bibliotecas: **pandas**, **numpy**, **sklearn**, e ferramentas de upload do Google Colab. Os dados foram carregados a partir do arquivo *games.csv*.

2.2 Análise inicial do dataset

Foi realizada uma análise exploratória da estrutura do dataset e visualização das primeiras linhas. Em seguida, calculou-se quais campeões foram mais selecionados. O top 10 mais frequente incluiu ids como **412**, **18**, **67**, entre outros.

2.3 Verificação de valores nulos, negativos e inválidos

Foram testadas condições básicas de validade:

- **gameId** $\leq 0 \rightarrow$ **0 ocorrências**
- **gameId**, **creationTime**, **gameDuration** $\rightarrow$ **0 valores nulos**

A integridade estrutural também foi avaliada:

- Cada partida deve possuir **10 campeões** $\rightarrow$ **0 inconsistências**
- Cada partida deve possuir no máximo **10 bans** $\rightarrow$ **0 inconsistências**

2.4 Validação de consistência

Foram aplicadas regras lógicas relacionadas ao jogo:

1. **gameDuration ≥ 300s** (evitando remakes)
    - 1195 partidas foram removidas por serem menores que 300s.
2. O time vencedor deve ter destruído ao menos **1 torre**.
    - Time 1 com vitória e 0 torres destruídas $\rightarrow$ **631 casos**
    - Time 2 com vitória e 0 torres destruídas $\rightarrow$ **589 casos**
    - Todos esses casos foram removidos.

2.5 Remoção de duplicatas

Foram identificados **429 gameIds duplicados**, todos mantendo dados absolutamente idênticos.
A solução adotada foi manter a **primeira ocorrência** e excluir as demais utilizando drop_duplicates().

2.6 Seleção de variáveis

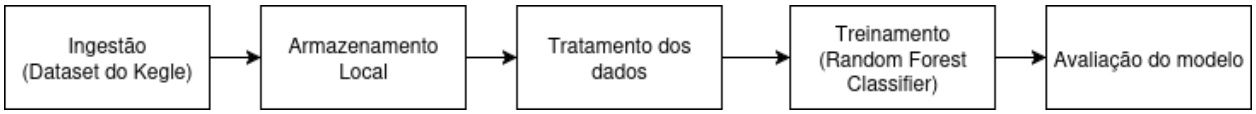
Para o modelo de previsão, o dataset foi reduzido às seguintes colunas:

- **t1_champ1id ... t1_champ5id**
- **t2_champ1id ... t2_champ5id**
- **winner**

O objetivo era prever o vencedor usando exclusivamente os campeões escolhidos.

2.7 Preparação do modelo

- **X:** campeões selecionados
- **y:** variável target (time vencedor)
- Split dos dados: **70% treino / 30% teste**, estratificado.
- Modelo utilizado: **Random Forest**, 200 árvores, random_state=42.



3. RESULTADOS

As métricas principais foram:

- **Acurácia:** 0,5185
- **Precisão ponderada:** 0,5181
- **F1-Score ponderado:** 0,5175

O relatório de classificação mostrou:

Relatório de classificação:

	precision	recall	f1-score	support
1	0.52	0.56	0.54	7823
2	0.51	0.47	0.49	7624
accuracy			0.52	15447
macro avg	0.52	0.52	0.52	15447
weighted avg	0.52	0.52	0.52	15447

Observa-se uma **leve vantagem** na predição para o time 1, mas próxima de aleatória.

O desempenho geral ficou em torno de **52%**, pouco superior ao puro chute (50%).

4. DISCUSSÃO

O desempenho modesto do modelo indica que **a escolha dos campeões por si só não é suficiente** para prever o resultado de uma partida ranqueada de League of Legends.

Outros fatores que influenciam diretamente o resultado não foram incluídos, como:

- Habilidade individual dos jogadores;
- Sinergia entre campeões (combinações específicas);
- Estatísticas in-game: ouro total, quantidade de abates, controle de objetivos;
- Estratégia de composição (early, mid, late game);
- Patch e meta do jogo.

Além disso, mesmo após limpeza rigorosa, o dataset apresenta partidas onde fatores externos ou não modelados influenciam fortemente o desfecho, reduzindo a capacidade preditiva do modelo.

5. CONCLUSÃO

O presente trabalho executou um processo completo de análise e preparação de um dataset de partidas ranqueadas de *League of Legends*, incluindo:

- Validação de integridade;
- Correções de inconsistências;
- Remoção de outliers e duplicatas;
- Seleção de variáveis;
- Aplicação de modelo preditivo.

O modelo Random Forest alcançou aproximadamente **52% de acurácia**, revelando que os **campeões escolhidos não são suficientes para prever com precisão o vencedor** em competições de alto nível.