MVP — Classificação de Medalhistas em Natação (1912–2020)

Autora: Luiza Oliveira Lima

Objetivo

Construir um classificador para prever se um(a) atleta conquistará medalha em provas olímpicas de natação, a partir de atributos da prova e do país.

Dados & Preparação

Dataset público: Olympic Swimming History 1912–2020 (URL RAW GitHub). Alvo: Medalha = 1 se Rank ≤ 3; 0 caso contrário. Engenharia: Distance_value (ex.: 4x100 → 400). Pré-processamento: One■Hot para Stroke/Gender/Team e padronização para Year/Relay?/Distance_value. Split 70/30 estratificado. Seeds fixadas para reprodutibilidade.

Modelagem & Avaliação

Modelos: Regressão Logística (baseline), Random Forest, XGBoost. Otimização: RandomizedSearchCV (CV=5). Decisão de classificação com threshold varrido (0.20–0.80); threshold final = 0.40 para priorizar recall de medalhistas. Métricas: accuracy, precisão/recall/F1 (classe medalha), ROC■AUC.

Resultados & Insights

Modelo final: XGBoost (tuned) + threshold 0.40. Variáveis mais relevantes: Team (país, com destaque para EUA), Relay?, Stroke, Distance_value, Year. Robustez: remover Team derruba métricas; agrupar Team em Top■N vs OTHER reduz dependência, porém com perda de desempenho.

Reprodutibilidade

Notebook executável fim■a■fim (MVP_1_final.ipynb). Dados carregados via URL pública. Ambiente recomendado: Google Colab ou Python 3.10+ com numpy, pandas, scikit■learn, xgboost, matplotlib.