PROJETO MACHINE LEARNING

ANÁLISE DE DADOS

ARTILHEIROS NO FUTEBOL MUNDIAL

Comparação de algoritmos de classificação aplicados ao dataset Global Football Goalscorers

Disciplina: Machine Learning | **Autor:** Lucas Lima

DOMÍNIO E PROBLEMA

O projeto utiliza dados reais do futebol mundial para prever se um jogador é um artilheiro de alto desempenho, definido como **High Scorer** (acima do 75º percentil).

RELEVÂNCIA DO DOMÍNIO:

- Análise de performance esportiva
- Scouting e prospecção de talentos
- Métricas avançadas (Analytics)
- Modelagem estatística de desempenho

BASE DE DADOS

O projeto utiliza o dataset **Global Football Goalscorers**, extraído do Kaggle. Ele contém registros históricos abrangentes de artilheiros em competições internacionais, oferecendo volume suficiente para análises de complexidade média.

Arquivos Disponíveis

- goalscorers.csv (Main)
- shootouts.csv
- results.csv
- former_names.csv

Atributos Relevantes

- Gols Marcados
- Partidas Jogadas
- Temporada
- Idade do Jogador
- Competição

PREPARAÇÃO E TRANSFORMAÇÃO

Um pipeline robusto de **Engenharia de Dados** foi implementado em Python para garantir a qualidade e consistência das entradas para os modelos de classificação.

Limpeza e Normalização
Imputação (Mediana/Moda) e StandardScaler

Codificação e Seleção
One-Hot Encoding e remoção de colunas irrelevantes

Definição do Alvo e Divisão
Target High Scorer (>75%) e Split 80/20 estratificado

ALGORITMOS UTILIZADOS

1

LOGISTIC REGRESSION

Modelo linear clássico utilizado como **baseline** para comparação de desempenho. Destaca-se pela simplicidade e eficiência computacional.

- Alta interpretabilidade dos coeficientes
- Avaliado em pipeline integrado com pré-processamento
- Ideal para estabelecer uma linha base de performance

2

RANDOM FOREST

Algoritmo de ensemble baseado em **árvores de decisão**, capaz de modelar interações complexas e não lineares entre as variáveis.

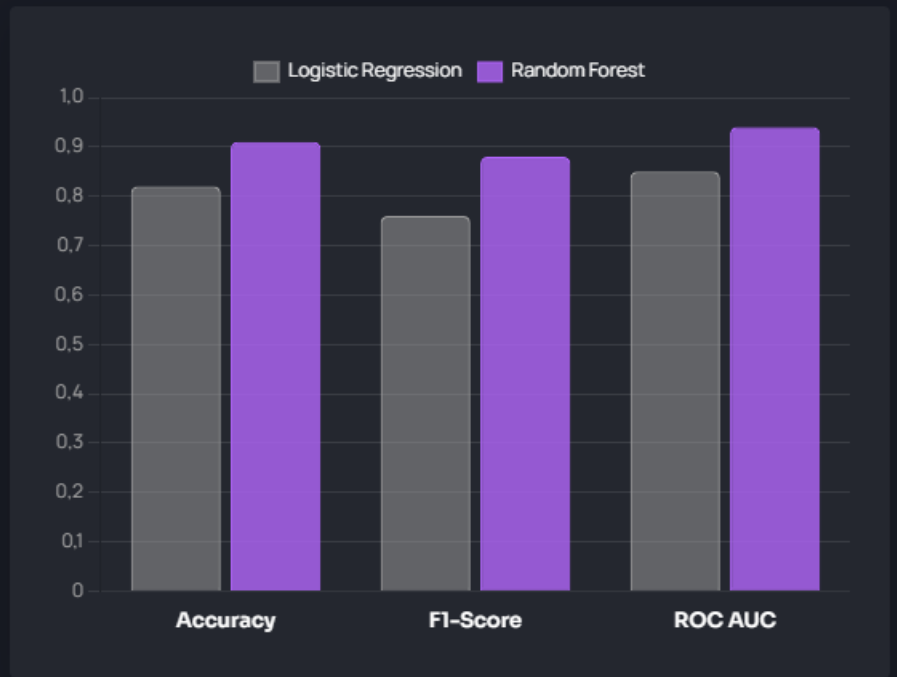
- Captura relações não lineares nos dados
- Fornece métricas de importância das features
- Geralmente apresenta performance superior em dados tabulares

RESULTADOS E AVALIAÇÃO

A comparação entre os algoritmos revelou uma clara superioridade do **Random Forest** em todas as métricas avaliadas, especialmente na capacidade de separação de classes (ROC AUC).

Features Mais Relevantes:

- Total de Gols na Temporada
- Partidas Jogadas
- Média de Gols por Minuto



CONCLUSÕES E REFERÊNCIAS

Este estudo confirmou a viabilidade do uso de Machine Learning para prever artilheiros de alto desempenho. A abordagem estatística e o uso de algoritmos robustos fornecem ferramentas valiosas para a modernização do scouting e análise esportiva.

Random Forest vs. Baseline

O Random Forest provou-se o modelo mais adequado para o problema, superando a Regressão Logística em métricas críticas como F1-Score e ROC AUC. Sua capacidade de lidar com não-linearidades foi decisiva para a classificação correta dos atletas.

Metodologia de Classificação

A definição de "High Scorer" baseada no percentil 75 mostrou-se eficaz. Este limiar permitiu uma separação de classes balanceada, capturando a elite dos artilheiros sem criar um desequilíbrio excessivo que prejudicasse o treinamento do modelo.

Pipeline e Referências

O pipeline desenvolvido assegura a reprodutibilidade e escalabilidade da análise. Dados obtidos do Kaggle (Global Football Goalscorers) e implementações via Scikit-Learn, servindo como base sólida para futuros projetos de analytics.