Identificação de jogadores atípicos no campeonato brasileiro de futebol com uso de aprendizado de máquina

Rodrigo M. A. Oliveira[1,2], **Ângelo M. O. Sant'Anna**[1,2], **Paulo H. Ferreira**[3]

[1]Programa de Pós-Graduação em Engenharia Industrial
[2]Escola Politécnica, Universidade Federal da Bahia
[2]Instituto de Matemática e Estatística, Universidade Federal da Bahia
Salvador - BA, Brasil

rodrigomarcel@ufba.br, angelo.santanna@ufba.br, paulohenri@ufba.br

Introdução

No futebol profissional, a avaliação de desempenho dos atletas é fundamental para tomadas de decisão que envolvem técnicas e estratégias. Geralmente, essas avaliações são feitas com base em indicadores isolados, como gols ou assistências. No entanto, essas métricas podem não capturar o desempenho global do jogador, especialmente considerando diferentes funções em campo. Este trabalho propõe uma abordagem baseada em detecção de *outliers*, utilizando o algoritmo *Isolation Forest* (iForest), para identificar jogadores com desempenhos atípicos, tanto positivos quanto negativos, ao longo das partidas. A técnica de *Análise de Componentes Principais* (PCA) foi empregada para reduzir a dimensionalidade dos dados e permitir a visualização dos jogadores em um espaço bidimensional, facilitando a identificação de padrões. O método *SHapley Additive exPlanations* (SHAP) foi utilizado para compreender quais variáveis mais influenciam a classificação de um jogador como fora do padrão. Os resultados permitem identificar de forma transparente os principais fatores técnicos e táticos associados aos desempenhos dos jogadores.

Principais objetivos

O presente estudo tem como objetivos: (i) classificar jogadores em grupos de alto desempenho, baixo desempenho e desempenho normal, ao longo das partidas; (ii) analisar os indicadores que mais contribuem para que um jogador seja considerado outlier; e (iii) gerar indicadores que possam servir de suporte à decisão para treinadores, analistas e departamentos de desempenho.

Estudo de caso

A base de dados utilizada compreende estatísticas individuais de jogadores do campeonato brasileiro de futebol masculino, que corresponde à Série A da temporada 2024. As variáveis utilizadas estão diretamente relacionadas ao desempenho técnico e tático dos jogadores, tais como: minutos jogados (Min.), gols marcados (Gols), assistências (Assis.), expectativa de gols (xG), expectativa de gols sem pênaltis (npxG), expectativa de assistências (xAG), ações que levam a finalizações (SCA), ações que levam a gols (GCA), precisão de passes (Cmp, Att, Cmp%), conduções de bola e dribles (Conduções, PrgC, Tent, Suc), bem como métricas associadas à construção de jogadas (PrgP). As informações foram extraídas da seguinte base de dados disponível no *Kaggle*: `brasileiro-player-stats-2024`.

Metodologia

O iForest, proposto por [2], é um modelo não supervisionado para detecção de *outliers* baseado em árvores que visa isolar amostras através de particionamentos recursivos. Para cada ponto inédito, o algoritmo gera uma pontuação, representada pela Equação (1), em que $E(h(x))$ é o valor médio das profundidades que um ponto inédito atinge em todas as árvores da floresta. O fator de normalização $c(k)$ (Equação (2)) representa a profundidade média mal-sucedida em uma busca na árvore binária, com k representando o número de pontos usados na construção da árvore; e a função $H(i)$ (Equação (3)) o número harmônico estimado.

$$S(x,k) = 2^{-\frac{E(h(x))}{c(x)}}, \tag{1}$$

$$c(x) = 2H(k-1) - \left(\frac{2(k-1)}{k}\right), \tag{2}$$

$$H(i) = \ln i + 0{,}5772156649. \tag{3}$$

O PCA é uma técnica de redução de dimensionalidade que transforma variáveis correlacionadas em um novo conjunto de variáveis ortogonais, denominadas componentes principais [1]. Essas componentes são combinações lineares das variáveis originais e preservam a maior variabilidade possível dos dados. No contexto deste trabalho, a PCA permite visualizar os padrões de desempenho dos jogadores em duas dimensões, facilitando a identificação de grupos e *outliers*. As variâncias explicadas por cada componente correspondem aos autovalores da matriz de covariância dos dados.

A metodologia SHAP, proposta por [3], é uma técnica utilizada para explicabilidade de modelos de aprendizado de máquina. Essa abordagem é baseada na Teoria dos Jogos e incorpora métodos de interpretações globais e locais como: importância de recursos, dependência de recursos e interações.

Resultados

A distribuição da idade dos jogadores está representada na Figura 1. Nota-se que a maior parte dos jogadores tem idade entre 20 e 35 anos. A Figura 2 representa a classificação dos jogadores para os 10 primeiros colocados em relação à quantidade de gols marcados, e a Figura 3 a quantidade de assistências por jogadores.

Os dados foram segmentados por posição, garantindo que a análise considerasse a natureza específica das funções desempenhadas pelos atletas dentro de campo. A análise descrita a seguir foi exemplificada com foco nos atacantes. Para garantir a comparabilidade entre variáveis que possuem escalas distintas, todos os atributos quantitativos foram padronizados utilizando a distribuição $N(0,1)$, ou seja, transforma as variáveis para uma distribuição com média zero e desvio padrão unitário.

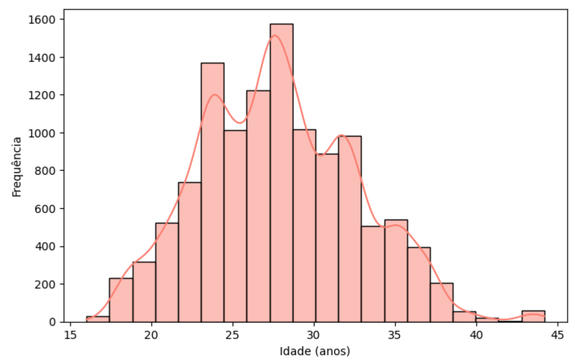

Figura 1: Distribuição da idade dos jogadores.

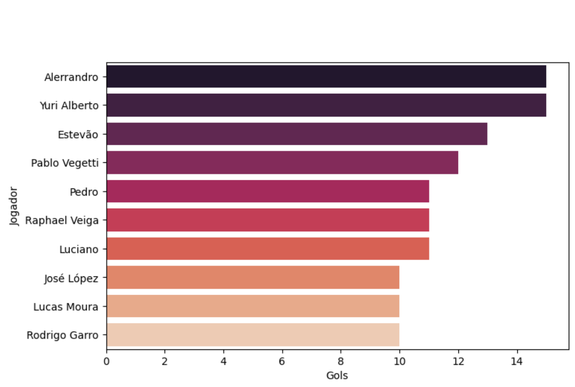

Figura 2: Classificação dos jogadores por gols.

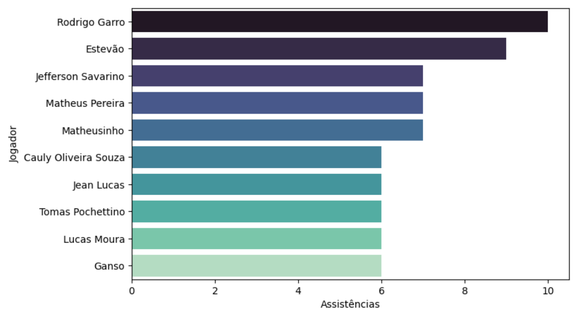

Figura 3: Classificação dos jogadores por assistências.

O iForest foi ajustado com um parâmetro de contaminação de 1%, refletindo a proporção esperada de jogadores considerados como desempenhos atípicos ao longo das partidas. Os valores negativos do *score* do modelo para cada jogador indicam instâncias normais e valores positivos indicam *outliers*. O resultado da classificação está disposto na Figura 4. Esse gráfico corresponde às informações das variáveis explicativas representadas em duas dimensões com auxílio da técnica PCA. A primeira componente representa 35,4% e a segunda 48,4% da variância total.

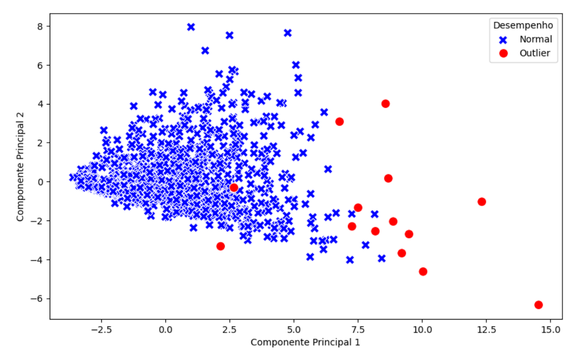

Figura 4: Distribuição dos dados em função do PCA e da classificação do iForest.

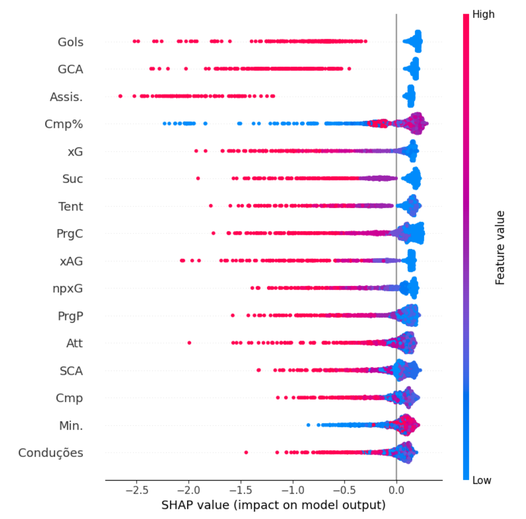

Figura 5: Explicabilidade global do modelo iForest.

A Figura 5 representa a explicabilidade global do modelo iForest com uso do SHAP. Ela fornece uma interpretação global dos resultados; para cada ponto no gráfico, representa uma observação da amostra. É possível identificar o impacto de cada variável para a decisão do modelo na detecção de *outliers*, conforme a tonalidade da cor. A quantidade de gols é a variável mais importante para discriminar jogadores atípicos na posição de ataque; portanto quanto maior é esse valor, maior a pontuação do iForest. Para a quantidade de minutos em campo, quanto menor esse valor, maior a pontuação.

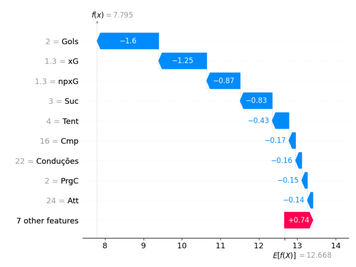

Figura 6: Outlier.

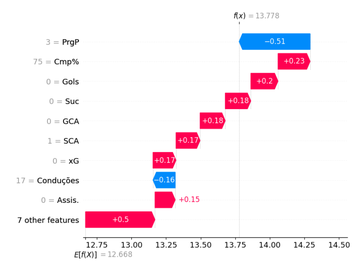

Figura 7: Normal.

As Figuras 6 e 7 representam amostras de jogadores classificados como *outlier* e normal, respectivamente. Elas fornecem uma explicabilidade local. Para o jogador classificado como *outlier* as variáveis com maiores pesos foram a quantidade de gols e expectativa de gols. Para o jogador classificado como normal as variáveis com maiores contribuições foram a precisão dos passes e quantidade de gols e dribles bem-sucedidos.

Os jogadores foram posteriormente classificados em três categorias distintas: (i) alto desempenho, representando os *outliers* de melhor performance; (ii) baixo desempenho, correspondendo aos *outliers* negativos; (iii) e desempenho normal, compreendendo os jogadores dentro do padrão estatístico da amostra. No contexto dos jogadores do ataque, a distinção entre alto e baixo desempenho levou em consideração a média da expectativa de gols. Portanto, jogadores com índice abaixo da média foram classificados como baixo desempenho. Isso resultou em 1346 partidas de jogadores classificados como normais, 12 com alto e 2 com baixo desempenho.

Conclusões

A abordagem com iForest foi capaz de identificar jogadores com desempenhos atípicos. A técnica SHAP permitiu explicar os fatores mais relevantes para essa classificação. Os resultados mostram o potencial do aprendizado de máquina como ferramenta de apoio à análise de desempenho no futebol.

Agradecimentos

Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Brasil - Código Financeiro 001

Referências

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning. 2009.

[2] Fei Tony Liu, Kai Ming Ting, and Zhi Hua Zhou. Isolation forest. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 413–422, 2008.

[3] Scott M Lundberg, Paul G Allen, and Su-In Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30, 2017.