Aplicação de algoritmos de aprendizado de máquina para predizer o desempenho de clubes de futebol europeus com base no comportamento no mercado e desempenho nas temporadas anteriores

Abel Gonçalves Chinaglia (ORCID: 0000-0002-6955-7187)[1,2] and Rafael Luiz Martins Monteiro (ORCID: 0000-0002-3208-6369)[1,2]

[1]Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Brasil
[2]Escola de Educação Física e Esportes de Ribeirão Preto, Universidade de São Paulo, Brasil

Abstract

O impacto financeiro das transferências de jogadores no futebol profissional tem crescido significativamente, o que reforça a necessidade de que os clubes realizem investimentos mais embasados em dados. Este trabalho propõe a aplicação de algoritmos de aprendizado de máquina supervisionado para predizer o desempenho futuro de clubes de futebol com base em suas atividades no mercado de transferências e no desempenho em temporadas anteriores. Serão utilizados dados das cinco principais ligas europeias — Bundesliga, Premier League, La Liga, Serie A e Ligue 1 — no período de 2009/2010 a 2019/2020. Serão construídas três bases de dados distintas: uma baseada apenas em dados de transferências, outra contendo métricas de desempenho histórico, e uma terceira combinando ambas. Os modelos buscarão classificar os clubes em três faixas de desempenho na temporada seguinte: parte superior (1º ao 6º lugar), intermediária (7º ao 14º) e inferior (15º ao 20º). Serão utilizados cinco algoritmos de classificação supervisionada — Random Forest, Regressão Logística, K-Nearest Neighbors, Support Vector Machine e Gaussian Naive Bayes — por meio da biblioteca Scikit-learn em Python. A avaliação será realizada com validação cruzada k-fold e otimização de hiperparâmetros via Grid Search. As métricas utilizadas incluirão acurácia, acurácia balanceada, precisão, revocação e F1-score. Espera-se observar diferenças na performance dos modelos entre as ligas e entre os diferentes conjuntos de dados. Além disso, espera-se que os modelos baseados apenas em desempenho histórico apresentem maior acurácia do que aqueles baseados apenas em dados de transferências, mas que a combinação dos dois conjuntos gere os melhores resultados. Também se espera que períodos mais longos de análise histórica resultem em modelos mais precisos. Este estudo pretende preencher uma lacuna na literatura ao propor a previsão do desempenho de clubes com base em comportamento de mercado, oferecendo ferramentas analíticas para apoiar decisões estratégicas no futebol profissional.

Keywords: aprendizado de máquina, futebol europeu, transferências de jogadores, previsão de desempenho, análise de dados esportivos

1 Introdução

As transferências de jogadores de futebol profissional podem ser determinantes para o destino do time na temporada e tem envolvido cada vez maiores valores financeiros [1]. O mercado de transferências movimenta quantias substanciais de capital, com taxas crescentes que impactam significativamente a saúde financeira dos clubes [2]. Em 2023, por exemplo, foram registradas 3.279 transferências com taxas, totalizando £7.6 bilhões, um aumento considerável em relação ao ano anterior [3].

Com o aumento do dinheiro sendo investido e o impacto no time, torna-se crucial gastar dinheiro da melhor forma possível. A inteligência artificial tem sido cada vez mais aplicada na análise do desempenho esportivo, para predição de desempenho e auxiliar em tomadas de decisão [4]. Estudos já foram realizados envolvendo o valor dos atletas no mercado, porém com uma abordagem de predição do valor dos jogadores com base em estatísticas dos mesmos[5]. Até o momento não foram encontrados artigos que buscaram predizer o desempenho do time em temporadas subsequentes com base nos dados acerca das ações daquele time no mercado, tais investigações podem ajudar a entender quais ações resultaram em melhora de

desempenho e auxiliar os clubes nas tomadas de decisão ainda mais envolvendo quantidades cada vez maiores de dinheiro investido.

2 Objetivos

2.1 Geral

Aplicar algoritmos de aprendizado de máquina para predizer o desempenho futuro de times de futebol por meio da análise das transferências de jogadores e desempenho nas temporadas anteriores nas cinco principais ligas europeias (Bundesliga - Alemanha, Premier League - Inglaterra, La Liga - Espanha, Serie A - Itália e Ligue 1 - França) ao longo das temporadas de 2009/2010 a 2019/2020 utilizando diferentes filtros temporais e variáveis preditoras.

2.2 Específicos

- Comparar a capacidade de predição dos algoritmos com diferentes filtros temporais nas transferências e desempenho anterior dos times;
- Comparar a acurácia das predições utilizando diferentes variáveis preditoras, sendo elas: 1) dados de transferências de jogadores; 2) dados relativos ao desempenho do time nas temporadas anteriores; 3) dados de transferências de jogadores e relativos ao desempenho dos times nas temporadas anteriores;
- Comparar a acurácia das predições nas diferentes ligas europeias;

3 Materiais e Métodos

3.1 Dataset

O conjunto de dados que será utilizado neste trabalho é composto por dados de transferências de atletas e da classificação dos clubes das ligas ao longo de várias temporadas e foram extraídos de duas fontes principais. Os dados referentes às transferências de jogadores foram extraídos do repositório `github.com/ewenme/transfers` [6]. Já os dados referentes à classificação dos clubes, ao longo das temporadas de 1992/1993 a 2021/2022 da Bundesliga, La Liga, Serie A, Premier League e Ligue 1, foram obtidos manualmente no site `fbref.com` [7]. A extração manual dos dados de classificação foi realizada temporada a temporada para garantir conformidade com os termos de uso do `fbref.com`, evitando o uso de métodos automatizados.

O dataset está estruturado contendo como colunas os dados: nome do clube, nome do jogador, idade, posição, clube envolvido na negociação, valor da negociação, tipo da transferência, perído da transferência, valores ajustados, nome da liga, ano, temporada, posição, partidas jogadas, vitórias, empates, derrotas, gols à favor, gols contra, diferença de gols, pontos, pontos por partida, número de espectadores, artilheiro do time, goleiro do time, notas, gols esperados, gols esperados contra, diferença de gols esperados e diferença de gols esperados por 90 minutos. O dataset de cada liga possui as mesmas colunas indicadas mas número diferente de linhas, sendo que o dataset da Bundesliga apresenta 13446 linhas, o da La Liga apresenta 15140 linhas, o da Ligue 1 apresenta 15764 linhas, o da Premier League apresenta 22976 linhas e o da Serie A apresenta 27147 linhas.

3.2 Pré-processamento

A partir do banco de dados descrito acima será realizada uma filtragem das variáveis de interesse. Inicialmente será feito um corte temporal para selecionar somente as transferências e demais informações das temporadas de 2009/2010 a 2019/2020. Foi escolhida esta janela temporal para evitar dados referentes a pandemia de COVID-19 tendo em vista que este foi um fator de grande impacto sob as dinâmicas de mercado e desempenho dos jogadores de futebol. Também serão selecionadas somente as transferências que envolveram dinheiro, excluindo empréstimo gratuito, volta de empréstimo ou fim de contrato com o time anterior e assinatura de novo contrato. Serão criados 3 bancos de dados para treinamento e comparação dos modelos. Abaixo segue a descrição deles:

- **Banco de dados 1: Transferência de jogadores.**. Serão utilizadas como variáveis preditoras a idade dos atletas, a janela de transferência, o valor da contratação e a quantidade de atletas que

foram contratados e vendidos. Todas essas variáveis serão calculadas para 4 posições de atuação dos atletas (goleiros, defesa, meio de campo e ataque) e para a entrada e saída dos jogadores;

- **Banco de dados 2: Desempenho do time nas temporadas anteriores**. Serão utilizadas como variáveis preditoras a posição do time na(s) temporada(s) anteriore(s), quantidade de vitórias, empates, derrotas, gols feitos, gols sofridos, saldo de gols e pontos por partida;
- **Banco de dados 3: Transferência de jogadores e desempenho do time nas temporadas anteriores**. Serão utilizadas como variáveis preditoras do banco de dados 1 e 2.

Para todos os bancos de dados a variável alvo a ser predita será a posição final do time na temporada seguinte. Para isso serão criadas três categorias, os times que ficaram na parte de cima da tabela (1º ao 6º colocados), no meio (7º ao 14º colocados) e na parte inferior da tabela (15º aos 20º colocados). Também serão testados diferentes filtros temporais nesses bancos de dados, sendo eles: dados de 1 ano anterior para prever o próximo, dados de 3 anos anteriores e dados de 5 anos anteriores. Os dados serão padronizados utilizando a técnica StandardScaler para que a magnitude das variáveis de entrada não influencie no treinamento dos modelos.

3.3 Algoritmos de Aprendizado de Máquina

Serão utilizados modelos de aprendizado de máquina supervisionado para classificação. Para isso, será utilizada a biblioteca de aprendizado de máquina em Python, a Scikit-learn[8]. Será comparado o desempenho de 5 modelos de aprendizado de máquina utilizados para classificação: Random Forest, Logistic Regression, K-nearest neighbors, Support Vector Machine e Gaussian Naive Bayes.

O Grid Search será utilizado para otimizar os hiperparâmetros dos modelos, escolhendo a melhor combinação com base na acurácia balanceada visando um desempenho consistente entre as classes. Serão passados valores de hiperparâmetros escolhidos previamente pensando na relação custo computacional e desempenho. A validação cruzada será utilizada para avaliar o desempenho dos modelos, com k = 5. As métricas escolhidas para avaliação dos modelos serão acurácia, acurácia balanceada, precisão, revocação, e F1 Score. A média dessas métricas serão reportadas tanto para cada classe quanto para a média geral[9].

3.4 Análise estatística

Tanto as métricas descritas na seção avaliação dos modelos de aprendizado de máquina quanto os valores das variáveis cognitivas, técnicas, táticas e físicas serão descritos em média e desvio padrão e comparados entre as diferentes categorias. A normalidade e homogeneidade serão verificadas pelos testes de Shapiro-Wilk e Levene. Quando os dados apresentaram distribuição normal e homogênea, eles serão comparados utilizando a Anova one-way e o post-hoc de Tukey's HSD (Honestly Significant Difference). Quando a distribuição dos dados não for normal e homogênea, as métricas serão comparadas utilizando o teste de Kruskal-Wallis com post-hoc de Dunn com ajuste de Bonferroni. Em todos os casos o nível de significância considerado será de $p < 0{,}05$. As análises estatísticas e o processamento de dados, incluindo treinamento dos modelos, será realizado por meio de algoritmos em Python 3.

3.5 Disponibilidade de dados

Todo projeto será realizado com controle de versão no Github e ao final será disponibilizado os códigos e dados.

4 Resultados esperados

Espera-se que os algoritmos apresentem um desempenho diferente nas diferentes ligas europeias, porém sem indicativo de qual seria melhor com base na falta de estudos deste tipo na literatura. Espera-se também que o desempenho nas temporadas anteriores apresentem melhores resultados quando comparados aos algoritmos com somente os dados de transferências, mas que ambos dados juntos apresentem resultados superiores. Como este tipo de investigação ainda não foi feito na literatura, predizendo desempenho futuro com base em dados de mercado, não temos um indicativo de qual algoritmo apresentará melhor performance. Por fim, espera-se que os algoritmos com maior corte temporal apresentem melhores resultados em relação aos demais.

References

[1] European Club Association (ECA), PricewaterhouseCoopers (PwC), LIUC Università Cattaneo (represented by Emanuele Grasso and Ernesto Paolillo). STUDY ON THE TRANSFER SYSTEM IN EUROPE; 2013. Acessado em abril 8, 2025. Available from: https://www.ecaeurope.com/media/2731/eca-study-on-transfer-system-in-europe.pdf.

[2] Veliz E. Here We Go! Predicting Transfer Market Valuations of Premier League Footballers; 2025. Acessado em abril 8, 2025. https://medium.com/stanford-cs224w/here-we-go-predicting-transfer-market-valuations-of-premier-league-footballers-1774131946ff.

[3] FIFA. FIFA Global Transfer Report 2023. FIFA; 2023. Available from: https://digitalhub.fifa.com/m/114622e4e17cf6a8/original/FIFA-Global-Transfer-Report-2023.pdf.

[4] Claudino JG, Capanema DdO, de Souza TV, Serrão JC, Machado Pereira AC, Nassis GP. Current approaches to the use of artificial intelligence for injury risk assessment and performance prediction in team sports: a systematic review. Sports medicine-open. 2019;5:1-12.

[5] Shen Q. Predicting the value of football players: machine learning techniques and sensitivity analysis based on FIFA and real-world statistical datasets. Applied Intelligence. 2025;55:265. Available from: https://doi.org/10.1007/s10489-024-06189-0.

[6] ewenme. transfers: data on European football player transfers; 2025. Repositório sem licença explícita; acesso em 10 nov. 2023. https://github.com/ewenme/transfers.

[7] Sports Reference LLC. FBref.com: Football Statistics and History; 2018. Acesso em 10 nov. 2023; uso automatizado proibido, extração manual permitida (cláusula 5) :contentReference[oaicite:2]index=2. https://fbref.com.

[8] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.

[9] Raschka S, Mirjalili V. Python machine learning: Machine learning and deep learning with python. Scikit-Learn, and TensorFlow Second edition ed. 2017;3:17.