Previsão de Sucesso de Atletas Jovens de Futebol usando Integração de diferentes Base de Dados

Lucas Calmon[1], Rodrigo Ferro[1], Carlos Pereira[1], Caio Santos[1], Lucas Giusti[1], Glauco Amorim[1], Eduardo Ogasawara[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)
20271-110 – Rio de Janeiro – RJ – Brazil

{lucas.calmon, rodrigo.ferro, carlos.pereira.1, caio.souza.4}@aluno.cefet-rj.br

{lucas.giusti, glauco.amorim}@eic.cefet-rj.br, eogasawara@ieee.org

Abstract. *There are various areas in soccer where prediction models can be utilized. Among them, identifying promising players can have a high cost-benefit ratio. Executive Functions (EF) are related to performance but have not yet been tested as predictors of success in soccer. This article investigates using EFs to select youth players using machine learning methods such as Logistic Regression, Naive Bayes, Decision Tree, and Random Forest to predict which players in the selected database were present and listed on the Transfermarkt platform. The best model was Random Forest combined with imputation, with a precision of 0.77. The present study indicates that EFs can be good predictors of success in soccer with up to 7 years of precedence.*

Resumo. *Há diversas áreas no futebol onde modelos de previsão podem ser utilizados, dentre elas, identificar jogadores promissores pode ter um alto custo-benefício. As Funções Executivas (FE) são relacionadas ao desempenho, mas ainda não foram testadas como preditores de sucesso no futebol. Este artigo investiga o uso de FEs para a seleção de jogadores da base com métodos de aprendizado de máquina como a Regressão Logística, Naive Bayes, Decision Tree e Random Forest para prever quais jogadores da base de dados estudada estavam presentes em uma plataforma confiável de dados: Transfermarkt. O melhor modelo foi o Random Forest combinado com imputação, com 0,77 de precisão. O presente estudo indica que as FEs podem ser bons preditores de sucesso no futebol com até 7 anos de antecedência.*

1. Introdução

Modelos de previsão estão se tornando cada vez mais integrados em diversas áreas de nosso cotidiano, incluindo o futebol. Nos clubes de futebol, essas ferramentas têm o potencial de revolucionar o desenvolvimento de atletas, especialmente os talentos das categorias de base. A previsão de lesões, por exemplo, é uma área emergente e pouco explorada pela comunidade de inteligência artificial e aprendizado de máquina [Beal et al., 2019]. Identificar jogadores com alta propensão a lesões permite à equipe técnica personalizar treinamentos e cuidados para minimizar riscos e otimizar o desempenho dos atletas. Outra aplicação promissora, mas ainda subexplorada, é a previsão da profissionalização de jovens jogadores.

As Funções Executivas (FEs) desempenham um papel crucial na avaliação e desenvolvimento de atletas nas categorias de base. Essas funções são essenciais para medir o desempenho de um jogador, pois influenciam diretamente a rapidez com que ele antecipa e se adapta às

situações em campo [Verburgh et al., 2016]. A capacidade de aprender habilidades motoras duradouras através da aprendizagem implícita e a tomada de decisões eficazes são significativamente impactadas pelas FEs [Verburgh et al., 2016].

As principais FEs – inibição, flexibilidade cognitiva e memória de trabalho – são fundamentais para um comportamento eficaz e direcionado a objetivos [Diamond, 2013]. Durante uma partida, os jogadores precisam constantemente avaliar suas opções e tomar decisões rápidas e precisas com base nas informações contextuais do jogo. A capacidade de um jogador de antecipar e reagir rapidamente às mudanças no posicionamento dos companheiros de equipe e adversários é essencial para seu sucesso.

Portanto, o objetivo desta pesquisa é desenvolver um modelo preditivo que possa antecipar a profissionalização de atletas das categorias de base dos clubes de futebol com anos de antecedência. Para isso, propomos um método que integra diversas fontes de dados e utiliza valores das FEs como base para o modelo preditivo.

Além desta introdução, o artigo está estruturado em quatro seções. A Seção 2 apresenta a fundamentação teórica. A Seção 3 detalha os métodos utilizados no estudo. A Seção 4 descreve os resultados obtidos. Finalmente, a Seção 5 discute as principais conclusões e sugere direções para pesquisas futuras.

2. Fundamentos

Estudos recentes têm explorado a correlação entre Funções Executivas (FEs) e o sucesso futuro dos atletas. Mello et al. [2021] demonstram que dados de FEs dos atletas, coletados entre 14 e 19 anos, estão relacionados à presença desses atletas na plataforma Transfermarkt[1] após os 23 anos de idade. Esse estudo destaca a importância de combinar FEs com o Efeito da Idade Relativa para prever o sucesso dos jogadores.

A previsão de desempenho de atletas utilizando aprendizado de máquina já tem sido explorada em diversos esportes. Soliman et al. [2017] demonstram a eficiência do algoritmo *Random Forest* na análise do desempenho de jogadores e na previsão da seleção dos melhores jogadores para o *All-Star Game*, avaliando a importância dos atributos para a precisão do modelo.

Van Bulck et al. [2023] investigaram a correlação significativa entre os resultados de jovens ciclistas sub-23 e seu sucesso futuro como profissionais. Utilizando um modelo de *Random Forest*, foram capazes de prever o potencial dos ciclistas baseando-se em resultados de competições juniores, idade, tipo de equipe contratante e desempenho em diferentes competições. Este modelo previu com sucesso a contratação dos 20 melhores ciclistas por uma equipe mundial.

Al-Asadi and Tasdemir [2022] utilizaram algoritmos de aprendizado de máquina para analisar dados de desempenho de jogadores de futebol e estimar seus valores de mercado. O estudo revelou que características como idade, desempenho esportivo e popularidade são influentes no valor de mercado dos jogadores, identificando quais dessas características são predominantes.

Apesar da vasta gama de estudos aplicando aprendizado de máquina em esportes para prever lucros, desempenho de jogadores e resultados de jogos, a aplicação desses conceitos no futebol ainda é limitada. da Silva Muniz and da Silva [2020] destacam a necessidade de otimização de receitas e controle de custos nos clubes de futebol. Investir nas categorias de base surge como uma alternativa importante, pois atletas formados pelo próprio clube representam menor custo salarial e maior potencial de receita futura em eventuais vendas.

[1] https://transfermarkt.com/

Com base nesses estudos, a presente pesquisa busca aplicar conceitos de aprendizado de máquina na previsão da profissionalização de jovens atletas de futebol, utilizando dados de FEs como variáveis preditivas. O modelo de aprendizado de máquina será avaliado pelos valores de acurácia, precisão, revocação e F1. Essa nova abordagem pode auxiliar os clubes de futebol, ajudando-os a tomar decisões mais informadas sobre o desenvolvimento de seus talentos.

3. Métodos

Esta seção descreve a abordagem metodológica utilizada no estudo. Primeiramente, é apresentado o processo de coleta de dados, detalhando as fontes e técnicas empregadas para colher as informações necessárias. Em seguida, são explicados os procedimentos de classificação, onde foram aplicados algoritmos de aprendizado de máquina para analisar e prever o sucesso dos atletas nas categorias de base.

3.1. Coleta de dados

Os dados deste estudo foram originalmente coletados nas categorias de base de três clubes de elite do Brasil, com a aprovação do comitê de ética da Universidade do Estado do Rio de Janeiro com o número de registro CRD42017077640. Os dados são compostos pelas funções executivas (FE) de cada jogador. A cronologia das bases está descrita na Figura 1.

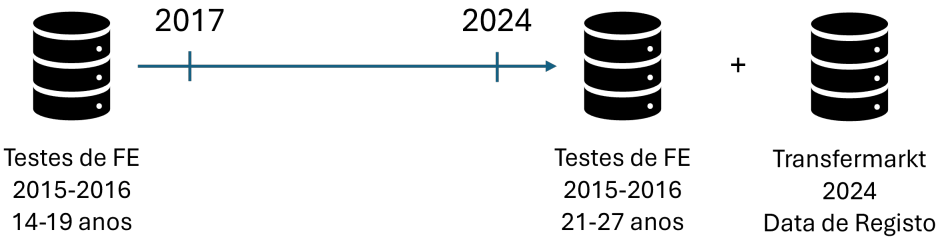

Figura 1. Cronologia da formação da base de dados utilizada para a análise

A base de dados consiste na junção de: (i) Testes de FE: Dataset das variáveis dos testes de FEs dos jogadores na base dos clubes, realizados entre 2015 e 2016; (ii) Alvo: Dataset das informações profissionais do atleta. A base de testes de FE contém os testes realizados, como o Teste de Trilhas (TMT), o *Color-Word Interference (Stroop)*, e o *Design Fluency Test* (DFT), descritos nessa seção, os quais foram utilizados para explorar a relação dessas habilidades com o potencial sucesso profissional dos atletas [Mello et al., 2021].

O TMT é composto por duas partes: Trilhas A e Trilhas B. A parte A é, principalmente, um teste de busca visual e velocidade motora em que as figuras são ordenadas como números, de 1 até 25, enquanto a parte B as figuras são ordenadas como números e letras subsequentemente (1, A, 2, B, 3, C), avaliando habilidades cognitivas de nível superior, como a flexibilidade mental. A diferença de tempo do teste A para o teste B resulta no TMTba [Shibuya-Tayoshi et al., 2007].

O teste de *Color-Word Interference*, mais conhecido como teste de *Stroop* consiste em três tarefas. A primeira é leitura de palavras, onde o participante lê em voz alta uma lista de palavras com nomes de cores em fonte preta no menor tempo possível. A segunda consiste na nomeação de cores das fontes de cada palavra que aparece. Na terceira tarefa, o participante tem que fazer a identificação das cores não correspondentes de fontes (palavra vermelho escrita na cor azul) e nomear a cor da fonte de cada palavra. O resultado foi considerado como o tempo necessário para responder o terceiro sub teste [Scarpina and Tagini, 2017].

O DFT avalia a capacidade do participante de gerar padrões geométricos não repetitivos com regras como usar um determinado número de linhas e pontos. Esse teste é realizado para avaliar a capacidade de planejamento e flexibilidade cognitiva com relação aos seu tempo de resposta [Chi et al., 2012].

A base de dados estudada contém 232 jogadores, todos com 21 anos ou mais no ano de 2024, cujos dados foram coletados do Transfermarkt, um site de estatísticas e transações de futebol amplamente utilizado como fonte de dados para estudos de futebol profissional [Bezuglov et al., 2023]. Cada jogador teve seu perfil verificado e atualizado perante o Transfermarkt. Aqueles que não possuem perfil no site aos 21 anos foram classificados como não tendo alcançado a posição de jogador profissional no teste.

A etapa de junção das bases citadas na Figura 1 envolve a inclusão do atributo alvo descrito acima na base de dados de Testes de FE. Em relação ao pré-processamento, alguns jogadores não realizaram determinados testes de FE na base, então dois métodos foram utilizados para lidar com esse problema: imputação (IM), onde valores faltantes desses jogadores foram substituídos pela média da coluna; e a remoção (DN) dos jogadores que não realizaram os testes de FEs utilizados na Classificação. Nota-se que esses valores faltantes podem ameaçar a validação do teste.

3.2. Classificação

Na última etapa, foram utilizados os métodos *Regressão Logística* (RL), *Naive Bayes* (NB), *Decision Tree classifier* (DT) e *Random Forest classifier* (RF) para classificar os atletas. Para analisar os resultados obtidos pelos classificadores, foi utilizada a matriz de confusão e as medidas de desempenho que dela derivam: Acurácia, Precisão, Revocação e F1-Score. Todas essas medidas são usadas para medir a qualidade dos modelos. As combinações de algoritmos preditores, variáveis e estratégia de imputação testados estão descritas na Tabela 1.

Tabela 1. Etapas da Análise

Ordem	Etapa	Variáveis	Utilização
1	Mann Whitney	Todas	NA
2	RL, NB, DT e RF	ST3 (S)	DN e IM
3	RL, NB, DT e RF	ST3, TMTba e DFT (STD)	DN e IM
4	RL, NB, DT e RF	ST3 e TMTba (ST)	DN e IM
5	RL, NB, DT e RF	ST3 e DFT (SD)	DN e IM

A base de dados foi particionada em 70% para o treinamento e 30% para o teste. Realizou-se o teste de *Mann Whitney* para identificar as variáveis que mostrassem diferença significativa entre os grupos de sucesso considerados. O valor alfa foi determinado em 0,05 para todos os testes. O resultado revelou que o único atributo significativo foi a terceira tarefa do teste de *Stroop* (ST3). Em seguida, para a utilização da base de dados, os testes foram realizados em diversas etapas: primeiro, aplicando apenas o ST3 como atributo previsor nas técnicas DN e IM (S); depois, utilizando ST3, TMTba e DFT (STD) como atributos previsores nas técnicas DN e IM; em seguida, realizando o teste com ST3 e TMTba (ST) como atributos previsores nas técnicas DN e IM; e, por fim, utilizando ST3 e DFT (SD) como atributos previsores nas técnicas DN e IM. As etapas 4 e 5 foram incluídas como testes de *ablação*, para entender a importância das variáveis DFT e TMTba respectivamente removidas. Todas essas etapas estão resumidas na Tabela 1.

4. Resultados

A Tabela 2 apresenta as métricas de desempenho de cada classificador na base de teste para cada modelo e a combinação de valores em cada classificador. Os resultados das métricas de avaliação mostram uma variação entre os diferentes métodos de classificação e utilização. O método NB apresentou uma Acurácia estável em todas as combinações de atributos preditivos, variando entre 0,66 e 0,70. Esse método também teve os melhores valores de Revocação assim como a RL, alcançando o valor máximo de 1 utilizando DN, indicando que ele é eficaz em identificar o máximo de positivos possíveis, o que pode ser considerado um fator importante no cenário das promessas do futebol. O NB também mostrou boas métricas de F1, chegando a 0,83.

O método de DT teve um bom valor de Acurácia, com 0,71 utilizando IM. Além disso, com a mesma utilização também alcançou uma Precisão de 0,77, o melhor nível de exatidão de todos os modelos testados. Apesar disso, o Revocação e F1-Score do modelo de DT foram menores em comparação com os outros.

O método de RF apresentou um desempenho variável, atingindo bons níveis de Acurácia juntamente com o DT. A precisão também foi elevada ao utilizar IM, alcançando 0,76. Esse método atingiu valores razoáveis em todas as métricas.

O método de RL teve o desempenho mais estável entre os modelos. Além de alcançar os melhores valores de Acurácia, Revocação (juntamente com NB) e F1, os valores das métricas de desempenho pouco variaram independente do número de variáveis preditivas utilizadas.

Tabela 2. Resultados das Métricas de Avaliação

Método	Utilização	Acurácia				Precisão			
		S	STD	ST	SD	S	STD	ST	SD
Naive Bayes	DN	0,70	0,70	0,66	0,70	0,71	0,72	0,70	0,72
	IM	0,70	0,67	0,67	0,70	0,72	0,71	0,71	0,72
Decision Trees	DN	0,60	0,55	0,54	0,51	0,70	0,72	0,71	0,68
	IM	0,71	0,62	0,60	0,61	**0,77**	0,74	0,73	0,72
Random Forests	DN	0,64	0,59	0,56	0,61	0,69	0,71	0,67	0,74
	IM	0,71	0,67	0,60	0,70	0,76	0,73	0,69	0,75
Logistic Regression	DN	0,71	**0,72**	0,71	**0,72**	0,71	0,72	0,71	0,72
	IM	0,70	0,70	0,70	0,70	0,72	0,72	0,72	0,72
Método	Utilização	Revocação				F1-Score			
		S	STD	ST	SD	S	STD	ST	SD
Naive Bayes	DN	**1.00**	0,97	0,94	0,97	0,83	0,82	0,80	0,82
	IM	0,96	0,92	0,92	0,96	0,82	0,80	0,80	0,82
Decision Trees	DN	0,76	0,62	0,59	0,62	0,73	0,67	0,65	0,65
	IM	0,86	0,76	0,71	0,76	0,81	0,75	0,72	0,74
Random Forests	DN	0,91	0,74	0,76	0,74	0,78	0,72	0,71	0,74
	IM	0,88	0,88	0,82	0,88	0,82	0,80	0,75	0,81
Logistic Regression	DN	**1.00**	**1.00**	**1.00**	**1.00**	0.83	**0.84**	0.83	**0.84**
	IM	0.96	0.96	0.96	0.96	0.82	0.82	0.82	0.82

Os melhores resultados foram atingidos ao utilizar apenas um atributo preditivo, o ST3. Enquanto os melhores valores de Precisão foram encontrados no método de DT utilizando IM, os melhores valores de Acurácia, Revocação e F1 foram atingidos com RL utilizando DN, evidenciando as peculiaridades de cada modelo.

Tais resultados podem ser comparados aos principais métodos atuais de seleção de talentos, que utilizam apenas a opinião subjetiva dos treinadores dos clubes. Werneck and Figueiredo [2024] mostra que apenas 15,9% dos jogadores da base dos clubes estudados alcançaram o Campeonato Brasileiro sub-20 ou o time profissional quando avaliados utilizando apenas a visão dos treinadores. Além disso, a antecedência de até 8 anos na previsão de profissionalização do atleta também é um ponto relevante da solução apresentada.

Enquanto todos os métodos são relativamente bons em prever o sucesso de jogadores, alguns modelos tiveram melhores resultados em características diferentes. O método DT com IM pode ser utilizado para investir melhor em certos atletas da base, graças à alta Precisão que proporciona uma menor quantidade de falsos positivos. Já o método de RL com DN pode ser utilizado em uma seleção de jogadores para capturar o máximo de casos positivos possíveis, graças ao alto valor de Revocação.

Em relação às etapas de ablação (4 e 5), é possível ver na Tabela 2 que quase todas as métricas apresentaram redução do desempenho quando comparadas ao modelo completo (STD). Especificamente, na comparação entre STD e ST (retirando DFT) somente o Revocação no modelo de Random Forests DN apresentou alguma melhora no desempenho. Fora esse caso, todos apresentaram queda do desempenho ou mantiveram os valores do STD. Tais resultados indicam que a variável DFT pode ser relevante, apesar das variações serem baixas.

Na comparação entre STD e SD (retirando TMTba), o desempenho foi melhor em SD geralmente, com algumas exceções específicas como no modelo de Decision Trees DN e IM para Acurácia, Precisão e F1-Score. Utilizando Regressão Logística não houve diferença nos valores entre STD e SD. Tais resultados indicam que, considerando um modelo que usa ST3 e DFT, incluir o TMTba pode ser pouco útil para previsão de sucesso no futebol.

5. Conclusão

Este artigo apresenta um estudo sobre a aplicação de modelos de previsão com o intuito de prever a profissionalização de atletas nas categorias de base. Os resultados obtidos utilizando apenas o ST3 como atributo apresentaram os melhores resultados de Precisão e Revocação. Este resultado mostra a relevância do Teste de Stroop na predição da profissionalização.

Este trabalho aponta o potencial de modelos de aprendizado de máquina baseados em FE na previsão de profissionalização de atletas jovens de futebol com até 8 anos de antecedência. É importante ressaltar que, como qualquer ferramenta de Machine Learning, o principal uso do modelo é fornecer informações diferenciadas para auxiliar as comissões das categorias de base nas tomadas de decisão.

Como trabalhos futuros, o estudo avaliará o impacto de diferentes alvos no modelo. Alguns exemplos são o valor de mercado de cada jogador aos 21 anos ou a idade do primeiro contrato profissional.

Referências

Al-Asadi, M. A. and Tasdemir, S. (2022). Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques. *IEEE Access*, 10:22631 – 22645.

Beal, R., Norman, T. J., and Ramchurn, S. D. (2019). Artificial intelligence for team sports: a survey. *Knowledge Engineering Review*, 34.

Bezuglov, E., Morgans, R., Butovskiy, M., Emanov, A., Shagiakhmetova, L., Pirmakhanov, B., Waśkiewicz, Z., and Lazarev, A. (2023). The relative age effect is widespread among European adult professional soccer players but does not affect their market value. *PLoS ONE*, 18(3 March).

Chi, Y. K., Kim, T. H., Han, J. W., Lee, S. B., Park, J. H., Lee, J. J., Youn, J. C., Jhoo, J. H., Lee, D. Y., and Kim, K. W. (2012). Impaired design fluency is a marker of pathological cognitive aging; results from the Korean longitudinal study on health and aging. *Psychiatry Investigation*, 9(1):59 – 64.

da Silva Muniz, L. and da Silva, M. (2020). Análise das demonstrações contábeis dos clubes brasileiros de futebol: comparação entre a situação econômica e financeira e o aproveitamento nas partidas oficiais de 2015 a 2017. *CAFI*, 3(1):17–32.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, 64:135 – 168.

Mello, M., Belloni, V., Vasconcellos, F., Soares, J., Ogasawara, E., and Giusti, L. (2021). Funções Executivas e Idade Relativa como Preditores de Sucesso no Futebol. In *Anais da Escola Regional de Informática do Rio de Janeiro (ERI-RJ)*, pages 111–118. SBC.

Scarpina, F. and Tagini, S. (2017). The stroop color and word test. *Frontiers in Psychology*, 8(APR).

Shibuya-Tayoshi, S., Sumitani, S., Kikuchi, K., Tanaka, T., Tayoshi, S., Ueno, S.-I., and Ohmori, T. (2007). Activation of the prefrontal cortex during the Trail-Making Test detected with multichannel near-infrared spectroscopy. *Psychiatry and Clinical Neurosciences*, 61(6):616 – 621.

Soliman, G., El-Nabawy, A., Misbah, A., and Eldawlatly, S. (2017). Predicting all star player in the national basketball association using random forest. In *2017 Intelligent Systems Conference, IntelliSys 2017*, volume 2018-January, pages 706 – 713.

Van Bulck, D., Vande Weghe, A., and Goossens, D. (2023). Result-based talent identification in road cycling: discovering the next Eddy Merckx. *Annals of Operations Research*, 325(1):539 – 556.

Verburgh, L., Scherder, E., van Lange, P., and Oosterlaan, J. (2016). The key to success in elite athletes? Explicit and implicit motor learning in youth elite and non-elite soccer players. *Journal of Sports Sciences*, 34(18):1782 – 1790.

Werneck, R. and Figueiredo, A. (2024). Goldfit Soccer: A Multidimensional Model for Talent Identification of Young Soccer Players. *Research Quarterly for Exercise and Sport*, 0(0):1–15.