



UNIVERSIDADE FEDERAL DO MARANHÃO
Engenharia da Computação - ECP

Ítalo José Silva Reis

Predição do Efeito Idade Relativa nas Copas do Mundo

São Luís

2026

Ítalo José Silva Reis

Predição do Efeito Idade Relativa nas Copas do Mundo

Projeto de Trabalho de Conclusão de Curso
submetido à Coordenação do Curso de
Engenharia da Computação da Universidade
Federal do Maranhão como requisito à
obtenção do grau de Bacharelado em
Engenharia da Computação.

Trabalho aprovado em São Luís, 14 de julho de 2026:

**Prof. Dr. Paulo Rogério de Almeida
Ribeiro**
Orientador

Prof. Dr. Bruno Feres de Souza
Examinador

Prof. Dr. Davi Viana dos Santos
Examinador

São Luís
2026

Reis, Ítalo.

Predição do Efeito Idade Relativa na Copa do Mundo / Ítalo Reis. – 2026.

35 p.

Orientador(a): Paulo Ribeiro.

Monografia (Graduação) – Curso de Engenharia da Computação, Universidade Federal do Maranhão, São Luís, 2026.

1. Efeito Idade Relativa. 2. Aprendizado de Máquina. 3. Machine Learning. I. Ribeiro, Paulo. II. Título.

RESUMO

O agrupamento de atletas por idade tem sido utilizado em esportes competitivos desde a formação das equipes mais novas até a entrada na categoria profissional. Essa forma de classificação foi criada com o objetivo de diminuir o contraste que existe entre indivíduos com diferentes níveis de maturidade biológica nas fases de transição entre a infância e a adolescência. Ainda assim, essa solução não tem sido suficiente para reduzir o Efeito da Idade Relativa (EIR), que surge devido a diferença de idade cronológica dos jogadores. Mesmo em um período de um ano, a variação de habilidades físicas, fisiológicas e cognitivas são consideráveis para desbalancear as condições de seleção e ofuscar os melhores talentos de cada grupo. Diversas análises têm sido realizadas no futebol sobre competições em que a participação é restrita ao público infantil e a adolescentes, mas análises amplas sobre esse fenômeno no futebol profissional ainda são bastante recentes. Portanto, esse trabalho tem o objetivo de preencher essa lacuna realizando um estudo sobre os dados de 22 Copas do Mundo. Nele é realizada uma análise exploratória sobre os jogadores que mais participaram desses campeonatos e suas posições em campo. Por fim se é utilizado diferentes tecnologias de aprendizado de máquina para prever o EIR na Copa do Mundo de 2026. Os resultados obtidos possibilitaram a verificação do EIR dentre os jogadores que mais jogaram nas copas do mundo, no entanto ainda houve limitações quanto à capacidade de predição do EIR nas futuras seleções.

Palavras-chave: Efeito Idade Relativa. Copa do Mundo. Futebol Profissional. Aprendizado de Máquina.

Lista de ilustrações

Figura 1 – Percentual de Jogadores da Liga Nacional de Hockey de 1983 por trimestre	12
Figura 2 – Deslocamento no viés de seleção causado pelo EIR	13
Figura 3 – Exemplo de Árvore de Decisão	18
Figura 4 – Percentual de Jogadores nascido no primeiro semestre por time vs N ^o de Partidas	25
Figura 5 – Teste do Qui-quadrado	27
Figura 6 – Teste de Odds-ratio	28
Figura 7 – Matriz de confusão	30
Figura 8 – Importância de Features	31

Lista de tabelas

Tabela 1	–	Tabela de contingência: jogadores por trimestre de nascimento	15
Tabela 2	–	Participações de Pelé em Campeonatos do Mundo	24
Tabela 3	–	Dados agregados por seleção para análise do EIR	26
Tabela 4	–	Resultados de desequilíbrio por equipe (chi-square vs. odds_ratio) . . .	28
Tabela 5	–	Distribuição percentual de jogadores por trimestre	29
Tabela 6	–	Distribuição percentual por posição e trimestre	29
Tabela 7	–	Métricas de Performance dos Modelos	30
Tabela 8	–	Efeito EIR na edição de 2026 ordenado por qui-quadrado	31
Tabela 9	–	EIR no campeonato de 2026 ordenada por odds-ratio	32

Lista de abreviaturas e siglas

EIR	<i>Efeito Idade Relativa</i>
LSTM	<i>Long Short-Term Memory</i>
RNN	<i>Recurrent Neural Network</i>
RMSE	<i>Root Mean Squared Error</i>
MSE	<i>Mean Squared Error</i>
MAE	<i>Mean Absolute Error</i>
IDH	<i>Índice de Desenvolvimento Humano</i>
PIB	<i>Produto Interno Bruto</i>
FIFA	<i>Fédération Internationale de Football Association</i>

Sumário

1	INTRODUÇÃO	8
1.1	Contexto	8
1.2	Objetivos	9
1.2.1	Geral	9
1.2.2	Específicos	9
1.3	Justificativa	9
1.4	Trabalhos Relacionados	10
2	REFERENCIAL TEÓRICO	12
2.1	Efeito Idade Relativa	12
2.2	Testes Estatísticos	14
2.2.1	Qui-quadrado	14
2.2.2	Razão de Chances	15
2.3	Modelos de Predição	16
2.3.1	Regressão Logística	16
2.3.2	Random Forest	17
2.3.3	XGBoost	19
3	METODOLOGIA	21
3.1	Revisão e Fundamentação Teórica	21
3.2	População e Amostra	21
3.3	Análise Exploratória	22
3.4	Tratamento dos dados	23
3.5	Modelos de Aprendizado de Máquina	23
4	RESULTADOS	25
4.1	Análise Exploratória	25
4.2	Aprendizado de Máquina e Predição	29
5	CONCLUSÃO	33
	REFERÊNCIAS	34

1 INTRODUÇÃO

1.1 Contexto

Nos esportes competitivos, devido às diferenças nos níveis de desempenho com o avanço da idade, é comum organizar equipes e competições por faixas etárias. Essa estratégia tem como objetivo proporcionar condições de treinamento e competição mais justas. No entanto, existem estudos que questionam se esse é o melhor método de classificação para grupos (BARNSELEY, 1985).

No boxe, por exemplo, que é um esporte que agrupa os atletas não apenas pela idade, mas também pelo peso, DELORME (2014) avaliou se existe uma lacuna de desempenho entre atletas amadores e encontrou a ausência dessa diferença, concluindo que o agrupamento de boxeadores por peso pode contribuir para competições mais justas em esportes de combate. ALBUQUERQUE et al. (2015), por outro lado, constatou que no judô olímpico, também um esporte de combate e organizado por idade, existe um desequilíbrio nas competições entre atletas masculinos da categoria peso pesado. Dessa forma, isso suscita uma discussão mais ampla sobre o tema.

O Efeito da Idade Relativa (EIR), como tem sido denominado na literatura, propõe que jogadores nascidos em meses próximos ao primeiro e ao segundo trimestres do ano de seleção para a competição apresentam uma maior vantagem competitiva (HELSEN; STARKES; Van Winckel, 1998)(BARNSELEY, 1985)(SIMMONS; PAULL, 2001).

Alguns estudos sugerem que esse efeito pode até ser causado pela diferença na maturidade biológica que atletas jovens podem apresentar quando selecionados dentro de uma mesma faixa etária (MALINA et al., 2000). Em contraste com a idade cronológica, a maturidade biológica está relacionada à idade esquelética, que engloba o tamanho corporal, a força muscular e outras características físicas em determinada idade. Atletas jovens podem ser agrupados pela idade cronológica, mas estar em estágios muito distintos de maturidade biológica, o que proporciona vantagem aos jogadores com maturação precoce (MALINA et al., 2000).

Nesse contexto, jogadores que não se beneficiam do EIR tendem a receber menor apoio e acompanhamento técnico, o que leva ao abandono esportivo em estágios iniciais (HELSEN; STARKES; Van Winckel, 1998).

Essas condições injustas no esporte durante as competições de base levam à propagação do EIR também no futebol profissional (SALINERO et al., 2013). Esse efeito cascata é reforçado em clubes que não realizam reajustes significativos, mantendo nos elencos atletas que já faziam parte das equipes das categorias mais jovens (MUJICA et al., 2009).

Diversos estudos foram conduzidos no futebol profissional em ligas nacionais ou continentais com o objetivo de investigar o EIR (DUARTE, 2025)(VITORIA, 2025)(SAAVEDRA-GARCÍA et al., 2016)(YAGÜE et al., 2018)(PADRÓN-CABO et al., 2017). SAAVEDRA-GARCÍA et al. realizou uma pesquisa sobre o futebol espanhol, analisando dados de 2116 jogadores durante os anos de 2009 e 2010. De forma semelhante, SALINERO et al. (2013) conduziu um estudo sobre o futebol europeu com 2763 jogadores, também em um período de tempo limitado.

Entretanto, poucos estudos no futebol profissional utilizam de uma base tão ampla para analisar o Efeito da Idade Relativa nas Copas do Mundo, que abrangem um período substancial de tempo e múltiplos países. O vigente Trabalho de Conclusão de Curso busca dar continuidade a outro trabalho realizado pelo mesmo autor (REIS, 2023), mas com um foco maior na análises dos dados que foram obtidos no primeiro momento.

1.2 Objetivos

1.2.1 Geral

Analisar o Efeito Idade Relativa nas edições das Copas do Mundo e desenvolver classificadores binários para prever o EIR nos times da Copa de 2026.

1.2.2 Específicos

1. Realizar uma análise exploratória dos dados com o objetivo de identificar padrões e levantar hipóteses sobre o Efeito Idade Relativa nos times participantes das Copas do Mundo.
2. Testar e avaliar a eficácia de modelos de classificação binário para formar seleções de 2026
3. Analisar, em comparação com as copas anteriores, o EIR nas seleções de 2026.
4. Realizar a documentação do trabalho para uso acadêmico na área de Educação Física.

1.3 Justificativa

O estudo dirigido por esse Trabalho de Conclusão de Curso visa desenvolver um entendimento melhor sobre as dinâmicas do Efeito Idade Relativa no futebol profissional e, particularmente, nas Copas do Mundo. O entendimento desse fenômeno contribui para que técnicos e seleções criem regras mais justas para triagem de jogadores ao longo de suas carreiras. Ele também contribui significativamente para literatura no âmbito do futebol

profissional, visto que o EIR é menos estudado nesse contexto em comparação com o das categorias de base.

Ademais, o uso de técnicas mais avançadas como algoritmos de aprendizado de máquina é uma forma de avançar em um tema esportivo de forma interdisciplinar com tecnologias que por vezes são mais abordadas no contexto acadêmico da computação. Modelos de *machine learning* podem encontrar padrões mais complexos de se detectar em análises de dados e têm ganhado cada vez mais espaço dentro da literatura (SAAVEDRA-GARCÍA et al., 2019)(JAUHIAINEN et al., 2019)(KELLY et al., 2022).

Dessa forma, o presente trabalho visa preencher essa lacuna, realizando análises que possam ser utilizadas por acadêmicos de Educação Física no estudo do EIR em campeonatos profissionais. Muitos dos métodos estatísticos utilizados no trabalho, como qui-quadrado e razão de chances, já são bem conhecidos na área, porém se introduz novos conceitos quanto à aprendizado de máquina que podem servir tanto a estudantes que estão sendo introduzidos ao estudo do EIR quanto à pesquisadores que não tem familiaridade com essas tecnologias.

1.4 Trabalhos Relacionados

A interseção entre o Efeito Idade Relativa (EIR) e o aprendizado de máquina no contexto do futebol constitui uma linha de pesquisa ainda em consolidação, mas com contribuições crescentes e metodologicamente diversificadas. Os trabalhos a seguir são representativos dessa trajetória e guardam relação direta com a abordagem adotada neste estudo.

SAAVEDRA-GARCÍA et al. (2019) propuseram uma das primeiras abordagens estatisticamente sofisticadas para o estudo do EIR em competições da FIFA. Utilizando regressão logística aditiva aplicada a uma amostra de 21.639 jogadores de torneios entre 1908 e 2012, os autores superaram a limitação do teste qui-quadrado tradicional ao incorporar variáveis preditoras como categoria etária, sexo e evolução histórica do fenômeno. Os resultados confirmaram a sobre-representação consistente de jogadores nascidos no primeiro trimestre, com variações em função da competição e do período analisado (SAAVEDRA-GARCÍA et al., 2019). Esse trabalho é relevante para o presente estudo por demonstrar que métodos preditivos são capazes de capturar nuances do EIR que análises meramente descritivas não detectam.

No campo da identificação de talentos, JAUHIAINEN et al. (2019) abordaram o problema sob uma perspectiva de detecção de anomalias, tratando jogadores de elite como casos raros em um conjunto majoritariamente composto por não-elite. Aplicando uma *one-class support vector machine* (SVM) a dados de 951 jogadores juvenis finlandeses de 14 anos, os autores obtiveram área sob a curva ROC média de 0,73, indicando capacidade

preditiva moderada. O trabalho evidencia a importância de tratar o desequilíbrio de classes — problema estrutural em bases de dados esportivas — como parte integrante do desenho metodológico (JAUHIAINEN et al., 2019).

KELLY et al. (2022) conduziram uma investigação multidisciplinar em uma academia de futebol inglesa, combinando dados físicos, técnicos, psicológicos e sociológicos de 98 jogadores das categorias Sub-9 a Sub-16. Aplicando regressão LASSO com validação cruzada sobre 53 variáveis coletadas em duas temporadas, os autores identificaram que o trimestre de nascimento foi uma das variáveis com coeficiente não nulo mais expressivo para a predição das avaliações de desempenho dos jogadores — jogadores nascidos no primeiro trimestre receberam avaliações significativamente mais altas do que os nascidos no quarto trimestre. Os autores concluem que o EIR se manifesta de forma implícita nas decisões dos avaliadores, mesmo quando não é conscientemente considerado (KELLY et al., 2022). Essa conclusão corrobora a relevância de incluir o trimestre de nascimento como variável de entrada em modelos preditivos voltados ao EIR.

ALTMANN et al. (2024) avançaram nessa direção ao investigar, ao longo de sete temporadas, os fatores associados à seleção e deseleção de jogadores de uma academia profissional alemã (TSG 1899 Hoffenheim). Com uma amostra de 13.876 observações e utilizando árvores com *gradient boosting*, os autores identificaram que velocidade, mudança de direção, salto com contramovimento, reserva aeróbica e habilidades técnicas foram os atributos mais preditivos da permanência nas categorias de base (ALTMANN et al., 2024). Embora o EIR não seja o foco central do trabalho, o estudo demonstra que modelos baseados em *boosting* são especialmente adequados para capturar interações complexas entre atributos físicos e técnicos em contextos de seleção esportiva com grandes volumes de dados.

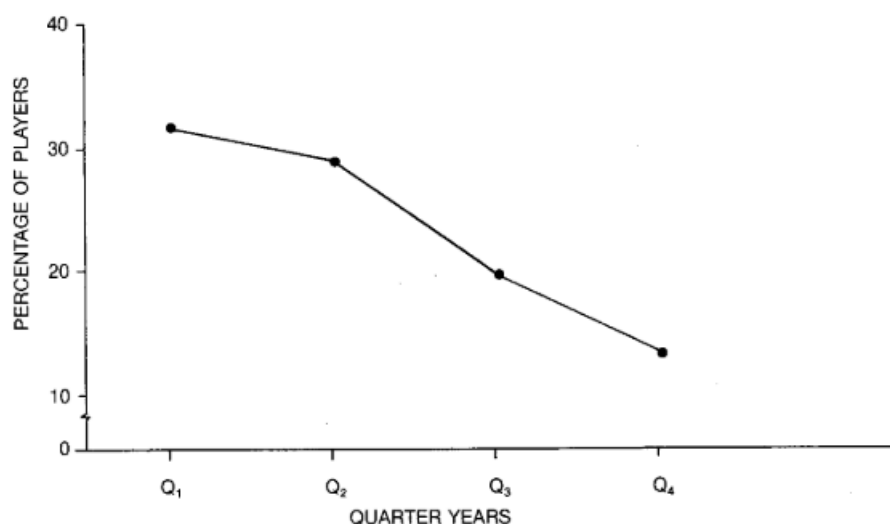
Em conjunto, esses trabalhos evidenciam uma convergência metodológica: a adoção de modelos preditivos supervisionados para investigar o EIR no futebol apresenta vantagens sobre as abordagens estatísticas tradicionais. O presente trabalho se insere nessa perspectiva ao empregar Regressão Logística, Random Forest e XGBoost para prever a ocorrência do EIR com base em variáveis de jogadores e contexto competitivo, contribuindo com uma abordagem orientada a dados aplicada ao futebol brasileiro.

2 REFERENCIAL TEÓRICO

2.1 Efeito Idade Relativa

O EIR é um fenômeno que ocorre em esportes nos quais a principal forma de agrupamento é pela idade que cada atleta detém em um intervalo pré-determinado de meses do ano (HELSEN; STARKES; Van Winckel, 1998)(BARNSELEY, 1985)(SIMMONS; PAULL, 2001). Um dos primeiros estudos desse efeito foi feito com o hockey em 1985. Barnsley (1985) categorizou jogadores de hockey de ligas nacionais e regionais dos Estados Unidos e Canadá por seus respectivos meses de nascimento. A classificação dos grupos na época eram feitas de 1º de Janeiro à 31 de Dezembro. Nesse contexto, pôde-se observar nos resultados uma predominância de jogadores que tinham seus meses de nascimento no início do ano de seleção, com a maior frequência de jogadores no mês de Janeiro conforme a Figura 1.

Figura 1 – Percentual de Jogadores da Liga Nacional de Hockey de 1983 por trimestre

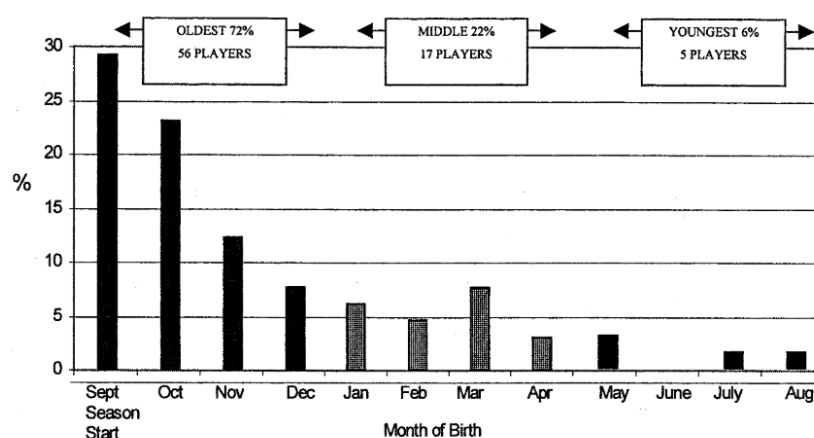


Fonte : (Barnsley, 1985)

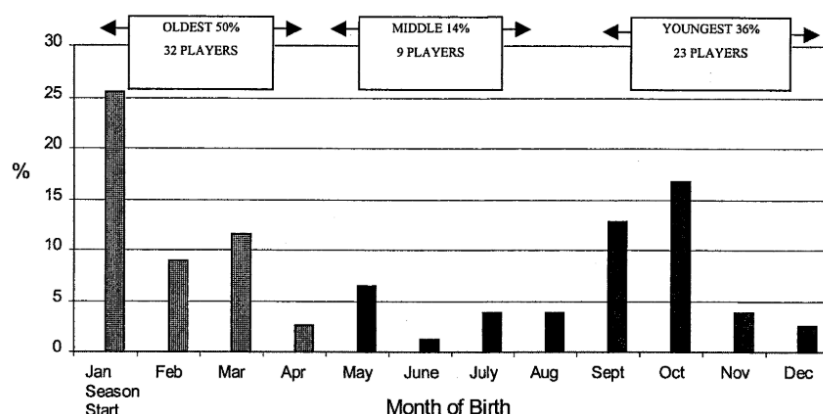
O estudo do EIR também tem sido feito sobre períodos de corte diferentes (HELSEN; STARKES; Van Winckel, 1998)(SIMMONS; PAULL, 2001). Helsen, Starkes e Van Winckel (1998) realizaram sua pesquisa com times de futebol da Bélgica, os quais tinham o ano de seleção iniciado em 1 de Agosto e terminado em 31 de Julho. Simmons e Paull 2001) compararam o ano de seleção para equipes que iniciaram em centro de treinamento profissional na Inglaterra (Figura 2a), com data de início em Setembro, e equipes que iniciaram na União das Associações Europeias de Futebol (UEFA) (Figura 2b), com início em Janeiro. Assim pôde-se concluir que o EIR não é exclusivo somente de uma única forma de categorização anual.

Figura 2 – Deslocamento no viés de seleção causado pelo EIR

(a) Distribuição de 78 jogadores escolares juniores da Inglaterra por data de nascimento



(b) Distribuição de 64 jogadores juniores da UEFA por data de nascimento



Fonte : (SIMMONS; PAULL, 2001)

Pesquisas tem sido realizadas, sobretudo, com equipes de jogadores juniores devido à prevalência do EIR nessa categoria. Autores como [Carli et al. \(2009\)](#) e [Mujika et al. \(2009\)](#) avaliaram em seus resultados que o EIR, como esperado, teve seu efeito mais atenuado em jogadores adultos que já participavam de equipes profissionais. Assim, crianças e adolescentes que estão nesse processo de iniciação de carreira como jogadores de futebol são os mais prejudicados por esse viés de seleção. Diferenças físicas e fisiológicas que surgem em períodos próximos à puberdade nesses jogadores contribuem para que eles tenham chances desiguais de competição dentro do mesmo grupo ([MALINA et al., 2000](#)), o que gera um alto índice de desistências e rejeição dos mais novos ([HELSEN; STARKES; Van Winckel, 1998](#)).

A presença do EIR em ligas profissionais resultante da perpetuação desse fenômeno ao longo dos campeonatos é chamada de efeito cascata ([MUJIKI et al., 2009](#)). Autores

como [Padrón-cabo et al. \(2017\)](#) tem investigado o EIR em ligas profissionais, mas estudos com ampla base de dados no futebol profissional ainda são recentes.

2.2 Testes Estatísticos

Uma das frentes para comprovar a presença de um fenômeno em uma determinada população é por meio de testes estatísticos. Um dos testes mais famosos e que serve de base para os demais é o teste da hipótese nula, que compara os dados observados com a distribuição esperada caso a hipótese seja verdadeira ([BUSSAB; MORETTIN, 2009](#)). O objetivo principal desse teste é confirmar se a hipótese é aceitável ou não, mas ainda sem mensurar a direção do fenômeno.

A depender da variável de estudo, pode se preferir determinados tipos de testes. No estudo do Efeito Idade Relativa, a qualidade do jogador ter nascido no primeiro, segundo, terceiro ou quarto trimestre do ano é uma variável categórica, pois expressa um atributo e não mera contagem ou mensuração ([BUSSAB; MORETTIN, 2009](#)). Dessa forma, dentre alguns testes para variáveis categóricas estão : Qui-quadrado e Razão de Chances (Odds Ratio).

2.2.1 Qui-quadrado

O Qui-quadrado é um teste que deriva da hipótese nula podendo ser calculado conforme a Equação 1. n_{ij} representa a frequência observada de um determinado evento ij .

$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \quad (1)$$

$$\mu_{ij} = n\pi \quad (2)$$

μ_{ij} é equivalente a expectativa do evento ij dado uma distribuição que segue a hipótese nula. Portanto, conforme a Equação 2, n é a frequência total e π é a probabilidade do resultado ocorrer por acaso ou conforme a hipótese nula. A diferença entre as duas variáveis mostra o quão distante ou próximo está o resultado observado para cada elemento ij da hipótese nula ([AGRESTI, 2007](#)).

Uma limitação importante do teste Qui-quadrado que pode ser notada na Equação 1 é que para valores pequenos de μ_{ij} o valor de X^2 tende a aumentar consideravelmente devido ao denominador da fração, o que pode resultar em uma impressão errada de que o resultado está bem mais distante da hipótese nula do que realmente está. Dessa forma, um requisito mínimo para realizar o teste Qui-quadrado sem adaptações é ter um $\mu_{ij} \geq 5$ ([AGRESTI, 2007](#)).

Para o Efeito Idade Relativa, em que a probabilidade da hipótese nula é 25%, visto que pode-se assumir que cada jogador tem aproximadamente 25% de probabilidade de nascer em cada trimestre do ano, obtém-se a seguinte condição pelo desenvolvimento da Inequação 3 :

$$\begin{aligned}\mu_{ij} &= 0.25 \cdot n \geq 5 \\ \Rightarrow n &\geq 20\end{aligned}\tag{3}$$

Portanto é necessário que cada time tenha pelo menos 20 jogadores para que seja aplicado o teste do qui-quadrado.

2.2.2 Razão de Chances

A Razão de Chances (*Odds Ratio*) é uma medida de associação utilizada para mensurar a magnitude e a direção de um fenômeno em determinada população. Ela pode ser calculada a partir de uma tabela de contingência 2x2, comparando a frequência de um evento entre dois grupos distintos. Se a chance (*odds*) de sucesso para o grupo experimental é $\frac{p_1}{1-p_1}$ e para o grupo de controle é $\frac{p_2}{1-p_2}$, em que p_x representa a probabilidade de sucesso no grupo x , a Razão de Chances é obtida pela divisão dessas duas quantidades conforme Equação 4:

$$\theta = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}\tag{4}$$

Quando $\theta > 1$, o grupo experimental apresenta maior chance de sucesso em relação ao grupo de controle (AGRESTI, 2007).

No contexto do Efeito Idade Relativa, a Razão de Chances permite comparar a frequência observada de jogadores nascidos no primeiro trimestre do ano com a frequência esperada sob uma distribuição uniforme. Para isso, organiza-se a Tabela 1, em que as linhas representam os trimestres de nascimento (Q1 vs. demais) e as colunas representam as frequências observadas e esperadas.

Tabela 1 – Tabela de contingência: jogadores por trimestre de nascimento

	Observado	Esperado
1º Trimestre	a	b
Demais Trimestres	c	d

Fonte : (Autor, 2026)

O cálculo de θ é realizado pelas Equações 5 e 6:

$$p_1 = \frac{a}{a+c}, \quad p_2 = \frac{b}{b+d}\tag{5}$$

$$\theta = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{\frac{a}{a+c} / \frac{c}{a+c}}{\frac{b}{b+d} / \frac{d}{b+d}} = \frac{a/c}{b/d} = \frac{ad}{bc} \quad (6)$$

Se $\theta > 1$, isso significa que a chance de um jogador ser nascido no primeiro trimestre na população observada é maior do que o esperado sob uma distribuição uniforme, indicando a presença do Efeito Idade Relativa.

2.3 Modelos de Predição

2.3.1 Regressão Logística

Modelos de regressão são utilizados para estudar o comportamento entre variáveis de resposta (dependentes) e variáveis explicativas (independentes). Sendo o objetivo principal dela encontrar o modelo no qual consegue descrever melhor a relação entre essas variáveis (HOSMER; LEMESHOW; STURDIVANT, 2013). Um dos modelos de regressão mais simples é o de regressão linear. O modelo de regressão linear pode ser descrito matematicamente pela Equação 7, que se assemelha a uma função afim. Nela, pode-se perceber que a média condicional da variável dependente $E[Y|X = x]$ pode variar de $-\infty$ a $+\infty$, uma vez que x também pode variar nesse intervalo.

$$E[Y|X = x] = \beta_0 + \beta_1 x \quad (7)$$

Ao contrário da regressão linear, a regressão logística é restrita a resultados binários e dicotômicos e a sua função se aproxima mais de uma função sigmoide, o que pode ser verificado na Equação 8. Ainda assim, o modelo de regressão linear é vantajoso devido a sua flexibilidade e facilidade de uso.

$$E[Y|X = x] = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (8)$$

Para verificar que o modelo é de fato restrito, ao contrário da regressão linear, pode-se testar os limites inferiores e superiores com $x \rightarrow \infty$ e $x \rightarrow -\infty$,

$$E[Y|X = x] = \lim_{x \rightarrow \infty} \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = 1$$

$$E[Y|X = x] = \lim_{x \rightarrow -\infty} \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{0}{1 + 0} = 0$$

Dessa forma é verificado que $0 \leq E[Y|X = x] \leq 1$.

Para se entender como um modelo de regressão logística é treinado é necessário entender como se estima seus parâmetros β_0 e β_1 . Similar à regressão linear, é utilizado o método de Estimativa de Máxima Verossimilhança. Esse método gera valores estimados para os parâmetros de forma a maximizar a probabilidade de se ter os dados observados (HOSMER; LEMESHOW; STURDIVANT, 2013).

No contexto do Efeito Idade Relativa, o problema de seleção de jogadores para os próximos campeonatos pode ser modelado como uma regressão com variável de saída binária Y , em que $Y = 1$ indica que o jogador foi convocado para a próxima Copa do Mundo e $Y = 0$ indica que não foi. Sendo Y uma variável de Bernoulli, sua esperança condicional é por uma soma de produtos que considera todas as possibilidades de Y e suas respectivas probabilidades como visto na Equação 9:

$$E[Y \mid X = x] = 1 \cdot P(Y = 1 \mid X = x) + 0 \cdot P(Y = 0 \mid X = x) = P(Y = 1 \mid X = x) \quad (9)$$

Ou seja, a esperança condicional de Y é simplesmente a probabilidade de o jogador ser convocado. Denotando $E[Y \mid X = x_i] = \pi(x_i)$ para simplificação, e assumindo independência entre as observações, a contribuição de cada jogador i para a função de verossimilhança é $\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$, de modo que a verossimilhança conjunta para uma amostra de n jogadores é dada por:

$$L(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (10)$$

Substitui-se a Equação 8 na 10 para se obter a função de verossimilhança que deve ser maximizada em função de β , que é o vetor de parâmetros o qual define $\pi(x_i)$ por meio da função logística.

$$l(\beta) = \prod_{i=1}^n \left(\frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right)^{y_i} \left[1 - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right]^{1-y_i}. \quad (11)$$

O modelo de regressão logística também possui limitações. Como visto anteriormente, o modelo pressupõe que as variáveis sejam independentes no cálculo da função de verossimilhança. Adicionalmente, é necessário uma quantidade considerável de observações por variável.

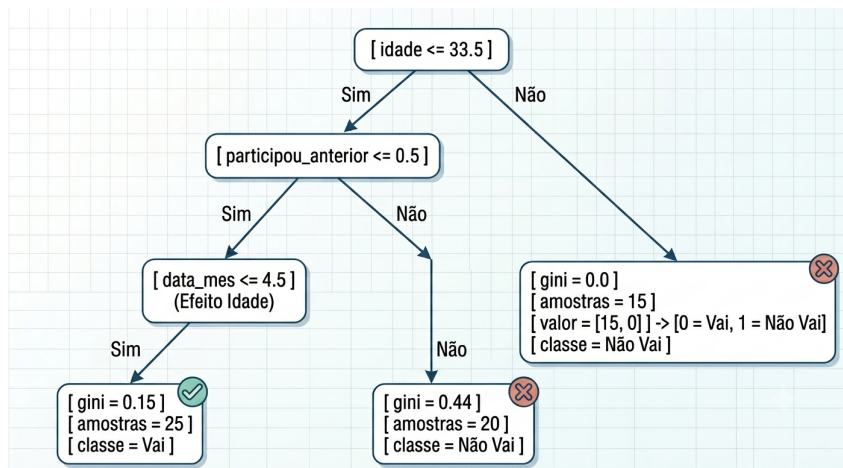
2.3.2 Random Forest

O *Random Forest* é um modelo de aprendizado que também pode executar atividades de classificação e regressão como o algoritmo de regressão logística. Um

diferencial é sua capacidade de aprender com *datasets* complexos e não lineares, assim como com variáveis não independentes. Para entender esse algoritmo é primeiro necessário introduzir conceitos como Árvore de Decisão, Modelos de aprendizado em conjunto (*Ensemble*), *Bagging* e *Pasting*.

A Árvore de Decisão treina a partir de variáveis independentes de forma a montar uma sequência de ramificações e isolar os grupos que estão ligados a cada uma das classes de resultados. Na Figura 3 temos um exemplo aplicado ao contexto do EIR. Nessa situação, a idade, participação na copa anterior e mês de nascimento são as variáveis explicativas. Ao organizar a idade como nó raiz, o algoritmo testa a possibilidade desse atributo ser um dos mais relevantes para determinar se o jogador participa ou não no próximo campeonato.

Figura 3 – Exemplo de Árvore de Decisão



Fonte : (Autor, 2026)

Se em uma amostra de 15 jogadores, todos com idade acima de 34 anos não participam do próximo campeonato então a árvore de decisão tem um bom desempenho e diz-se que tem um índice de gini pequeno indicando pureza, conforme o ramo direito da Figura 3. Se de 15 jogadores com idade acima de 34 anos, 7 participam e 8 não participam, então o algoritmo não conseguiu isolar uma variável independente que explique de forma satisfatória a variável resposta, portanto tendo um coeficiente de gini impuro (Géron, 2019). A fórmula da Equação 12 descreve o método de calcular esse atributo, sendo $p_{i,k}$ a razão entre o número de instâncias presente na classe k para uma instância de treino em i . Por exemplo, o número de jogadores que comparecerão ($k = \text{comparecem}$) no próximo campeonato sobre a quantidade total de jogadores que nasceram antes do mês de abril ($i = \text{nasceram antes do mês de abril}$).

$$G_i = 1 - \sum_{k=1}^n p_{i,k}^2 \quad (12)$$

A função custo para o treino de Árvore de Decisão é dada pela Equação 13. O algoritmo divide o dataset em dois usando uma variável explicativa k e um limiar t_k (ex : Idade do jogador ≤ 33 anos). Em seguida se busca os pares de k e t_k que produz os subconjuntos mais puros por meio da minimização da função custo.

$$J(k, t_k) = \frac{m_{left}}{m} G_{esquerda} + \frac{m_{right}}{m} G_{right} \quad (13)$$

Árvore de decisão é um modelo simples de fácil interpretação e versátil, porém é bastante sensível a pequenas variações no grupo de treino, o que motiva o uso de modelos de aprendizado em conjunto. Ao usar um modelos de *ensemble*, as predições de um grupo de classificadores ou regressores gera resultados de predição melhores que cada modelo individualmente (Géron, 2019).

Um dos métodos mais utilizados para se obter classificadores desse tipo de forma eficiente é por meio do *bagging* e o *pasting*. Nesses métodos, o conjunto de treino é dividido em subconjuntos de forma aleatória para treinar diferentes preditores do mesmo modelo. Para o *bagging* cada subconjunto retirado do principal é devolvido para o dataset antes de que seja treinado o próximo preditor. Para o *pasting* isso não acontece.

O *Random Forest* é um modelo de *ensemble* que é treinado geralmente com o método de *bagging* utilizando Árvore de Decisões. Ao invés de buscar pela melhor variável explicativa no momento de divisão de nós, o algoritmo busca a melhor variável explicativa dentro de um subgrupo aleatório de variáveis explicativas (Géron, 2019). No estudo de Ho (1995), o autor demonstra por meio da classificação de mais de 60000 amostras de dígitos manuscritos que a aplicação do modelo de Random Forest resulta em uma generalização melhor do que a de árvore de decisões sem comprometer o aprendizado no conjunto de treino e mantendo uma acurácia significativa.

2.3.3 XGBoost

Boosting é um método de *ensemble* que pode ser utilizado para diversos modelos de regressão e classificação. Diferente do *bagging*, que treina Árvores de Decisão em diferentes partes do conjunto de treino e depois forma um preditor composto por elas, para o *boosting* cada Árvore de Decisão é treinada sequencialmente em uma versão diferente do dataset (JAMES et al., 2021).

O AdaBoost (*Adaptive Boosting*), por exemplo, é um dos algoritmos que utilizam de *boosting* mais clássicos. Primeiramente um classificador como a Árvore de Decisão é treinada sobre o conjunto de treino. Em seguida, o peso sobre as instâncias que não foram classificadas corretamente aumenta. O próximo classificador é treinado com esses pesos,

o que força ele a ter uma atenção maior nas instâncias que foram classificadas de forma errada (Géron, 2019).

O *Gradient Boosting*, proposto por FRIEDMAN (2001) generaliza o *boosting* ao enquadrá-lo como um problema de otimização: o objetivo é construir, iterativamente, um modelo aditivo $F(x)$ que minimize uma função de perda L — como a log-verossimilhança para classificação ou o erro quadrático médio para regressão (FRIEDMAN, 2001).

A cada iteração m , uma nova árvore de decisão h_m é ajustada não sobre os rótulos originais, mas sobre os **resíduos** do modelo atual, isto é, a diferença entre o valor real e a predição corrente. Esses resíduos equivalem ao gradiente negativo da função de perda e indicam a direção na qual o modelo deve ser corrigido. O modelo é então atualizado de forma incremental:

$$F_m(x) = F_{m-1}(x) + \eta h_m(x), \quad (14)$$

onde $\eta \in (0, 1]$ é a taxa de aprendizado, que controla o quanto cada árvore contribui para o modelo final. Ao contrário do AdaBoost, que ajusta pesos por instância, o *Gradient Boosting* é mais flexível e suporta qualquer função de perda diferenciável (FRIEDMAN, 2001).

O *Extreme Gradient Boosting* (XGBoost), proposto por CHEN; GUESTRIN (2016), é uma implementação escalável e aprimorada do *Gradient Boosting* que adiciona regularização explícita à função objetivo, tornando o modelo mais resistente ao sobreajuste (CHEN; GUESTRIN, 2016). A função objetivo do XGBoost combina a perda de predição com um termo de penalidade sobre a complexidade de cada árvore adicionada.

3 METODOLOGIA

O objetivo desse projeto é realizar uma análise do Efeito Idade Relativa ao longo das Copas do Mundo utilizando técnicas estatísticas e de aprendizado de máquina. A metodologia utilizada segue as seguintes etapas fundamentais: Revisão e Fundamentação Teórica, Análise Exploratória dos Dados, Tratamento dos dados, Análise de Performance, Refinamento de Hiperparâmetros.

3.1 Revisão e Fundamentação Teórica

A revisão e fundamentação teórica para o projeto de análise do EIR envolveu o estudo de técnicas estatísticas e computacionais consolidadas na literatura sobre o tema. Por meio dessa atividade, foi possível obter uma base referencial sólida para a posterior comparação dos resultados obtidos com o panorama científico atual.

Durante o levantamento bibliográfico, foram exploradas técnicas estatísticas como o teste do Qui-Quadrado e a Razão de Chances (*Odds Ratio*). Essas abordagens justificam-se por servirem como indicadores robustos da existência de assimetria na distribuição das datas de nascimento, permitindo também mensurar a intensidade do fenômeno em diferentes campeonatos e clubes (AGRESTI, 2007).

Adicionalmente, no contexto de *Machine Learning*, buscou-se mapear o estado da arte para identificar se estudos prévios já utilizavam técnicas de aprendizado de máquina aplicadas ao EIR. Essa investigação foi crucial para compreender os principais ganhos e limitações de cada abordagem computacional, garantindo a extração de padrões relevantes e evitando o ajuste sobre ruídos.

Diante disso, no caso do teste do Qui-Quadrado, avaliou-se se o tamanho das classes atendia à condição teórica necessária de frequência esperada ($\mu_{ij} \geq 5$). Já na etapa de predição, analisou-se se o volume de dados histórico do número de Copas do Mundo disputadas por cada jogador (classes) era suficiente para o treinamento dos algoritmos (GÉRON, 2019), viabilizando a comparação de desempenho entre os modelos de Regressão Logística, *Random Forest* e *Gradient Boosting*.

3.2 População e Amostra

O presente estudo é delimitado pelo contexto do futebol profissional de elite. Ao contrário das categorias de base (*sub*), os campeonatos em foco compreendem as edições da Copa do Mundo da FIFA, nas quais os atletas apresentam uma maturidade biológica e idade cronológica mais avançadas. A população utilizada constitui-se dos jogadores que integraram os elencos oficiais de pelo menos uma edição do torneio entre os anos de 1930 e 2022.

Para a execução do trabalho, foram utilizados os dados históricos de 8.455 jogadores, obtidos a partir do mapeamento prévio realizado por (REIS, 2023). Como o próprio autor menciona, devido à antiguidade de algumas edições da competição, não foi possível encontrar registros completos para todas as Copas extraídas.

Diante disso, realizou-se um processo de tratamento e saneamento dessa base de dados para viabilizar sua posterior aplicação nos modelos de aprendizado de máquina. Esse procedimento consistiu na remoção de 4 seleções (totalizando 82 jogadores) que apresentavam dados substancialmente incompletos, além do preenchimento do mês de nascimento de 52 jogadores por meio da imputação pela mediana de seus respectivos times na edição correspondente da Copa do Mundo.

3.3 Análise Exploratória

A análise exploratória de dados constituiu uma das primeiras etapas executadas neste estudo. Por sua natureza, ela ocorre antes de qualquer procedimento de tratamento ou processamento avançado, tendo como objetivo principal desenvolver um entendimento aprofundado sobre o comportamento do conjunto de dados. Para tanto, buscou-se analisar as medidas descritivas de cada série temporal, as correlações entre as variáveis e o perfil das mudanças comportamentais entre as amostras (NIELSEN, 2021).

As medidas descritivas, como média, mediana, moda e desvio padrão, ajudaram a resumir o volume de dados coletados, facilitando a identificação de padrões, variações, sazonalidades e possíveis valores atípicos (*outliers*). Nesse projeto, tais métricas foram utilizadas como um diagnóstico inicial, considerado essencial antes da aplicação de métodos estatísticos inferenciais ou modelos de maior complexidade.

As correlações entre variáveis, por sua vez, mostraram-se importantes porque permitem identificar relações entre diferentes fenômenos, evidenciando se e como eles se movem conjuntamente ao longo do tempo. Essa compreensão auxiliou na explicação de comportamentos observados, na seleção de variáveis relevantes para análises e modelos preditivos, e na prevenção de interpretações equivocadas decorrentes de relações espúrias ou redundantes (NIELSEN, 2021).

Na prática, uma das principais frentes de análise exploratória consistiu em correlacionar a quantidade de edições de Copas do Mundo disputadas por cada seleção em relação à proporção de jogadores nascidos no primeiro semestre do ano civil. Adicionalmente, realizou-se um mapeamento das características dos atletas com maior longevidade no torneio, considerando aqueles que participaram de pelo menos quatro edições da competição. Ambos os estudos serviram de insumo fundamental para a modelagem e compreensão do EIR nas etapas subsequentes.

3.4 Tratamento dos dados

O tratamento de dados foi uma etapa fundamental dentro do projeto. Após a realização da análise exploratória, foi necessário retirar valores faltantes e normalizar os dados para que os modelos consigam atribuir pesos corretamente a cada variável do sistema.

Se os dados de treinamento contiverem muitos erros, valores discrepantes e ruído (por exemplo, decorrentes de medições de baixa qualidade), a identificação dos padrões pelo sistema será dificultada, reduzindo a probabilidade de um bom desempenho. Assim, foi importante investir tempo nesse processo.

No contexto do projeto, é preciso reconhecer que a dificuldade de se manter registros de times em campeonatos muitos antigos como há mais de 80 anos resulta em uma presença de datas de nascimento faltantes considerável (HEIM:SPIEL, 2000-2023). Times incompletos, por exemplo, podem conter mais jogadores na base de dados pertencente ao primeiro semestre e gerar um falso positivo, ou de forma contrária um falso negativo. Na base de dados foi verificado que 78.1% dos dados faltantes eram correspondente de times que estavam com até quatro jogadores sem datas de nascimento. Um percentual menor desses dados de copas muito antigas, como no caso da Indonésia na copa de 1938, tiveram mais de 5 jogadores com dados faltantes.

Para tratar essas situações, foi considerado técnicas de higienização dos dados como a remoção dos times de determinadas copas em que a quantidade de seus jogadores sem datas de nascimento superasse 5 jogadores. Por outro lado, foram imputados dados faltantes por meio da mediana da data de nascimento para os jogadores dos seleções que tiveram o valor inferior a essa quantidade. Dessa forma foram excluídos 4 times em copas específicas e imputado dados em 28 times em seus respectivos torneios.

Para realizar uma análise por time do EIR, realizou-se uma agregação dos times ao longo das Copas do Mundo. Dessa forma, foi calculado, por exemplo, o percentual médio de jogadores nascidos em cada trimestre e semestre ao longo das copas do mundo. Também foi calculado a quantidade de Copas do Mundo em que cada time participou e a quantidade de edições em que a seleção teve um EIR significativo, seja por odds ratio ou qui-quadrado. Essas informações foram úteis para análises de correlação que serão ilustradas melhor na seção de Resultados.

3.5 Modelos de Aprendizado de Máquina

O uso de aprendizado de máquina dentro do projeto foi realizado com o objetivo de analisar tendências do EIR para a Copa do Mundo de 2026. Com a quantidade de dados disponíveis das últimas 22 edições desses campeonatos e o avanço da tecnologia com relação a modelos de aprendizagem de máquina, o estudo de tendências do EIR para

os próximos anos se torna o tema discutível e relevante. Nesse trabalho, desenvolveu-se modelos de classificação com a intenção de prever se um jogador que já participou de uma edição da Copa seria eleito para o próximo campeonato e em cima desses resultados analisar tendências do EIR. O problema consiste em uma classificação binária, em que a classe "Participa" ou classe "um" se refere aos jogadores que são eleitos para a Copa do Mundo de 2026 e a classe "Não Participa" ou classe "zero" se refere aos jogadores que não serão eleitos para essa edição da copa.

Para se trabalhar com modelos supervisionados nesse cenário, cada jogador em seu respectivo campeonato tem a informação de se ele participou ou não da próxima Copa do Mundo conforme a Tabela 2 exemplifica. Nesse exemplo, na primeira linha da tabela, pode-se verificar não só que Pelé participou da edição de 1958, mas já se tem antecipadamente a informação de que ele participará da próxima copa de 1962 ($Label = 1$). Em 1970, no entanto, o *label* zero indica que não há uma participação em 1974. Portanto, o objetivo do algoritmo pode ser traduzido em prever, para cada jogador, o valor do label quando o ano do jogo é 2022 ainda que ele não tenha participado de fato nessa edição da Copa do Mundo.

Tabela 2 – Participações de Pelé em Campeonatos do Mundo

JOGADOR	TIME	ANO DE JOGO	FREQUÊNCIA	LABEL
Pelé	Brazil	1958	1	1
Pelé	Brazil	1962	2	1
Pelé	Brazil	1966	3	1
Pelé	Brazil	1970	4	0

Fonte : (Autor, 2026)

Para o treinamento, os dados foram divididos em conjunto de treino (80%) e conjunto de teste (20%) com estratificação pelo rótulo para garantir ambas as classes nos dois conjuntos. Também foi aplicada a técnica SMOTE (Synthetic Minority Over-sampling Technique) exclusivamente sobre o conjunto de treino. O SMOTE gera amostras sintéticas da classe minoritária por interpolação entre vizinhos mais próximos, evitando que o modelo seja tendencioso para a classe majoritária.

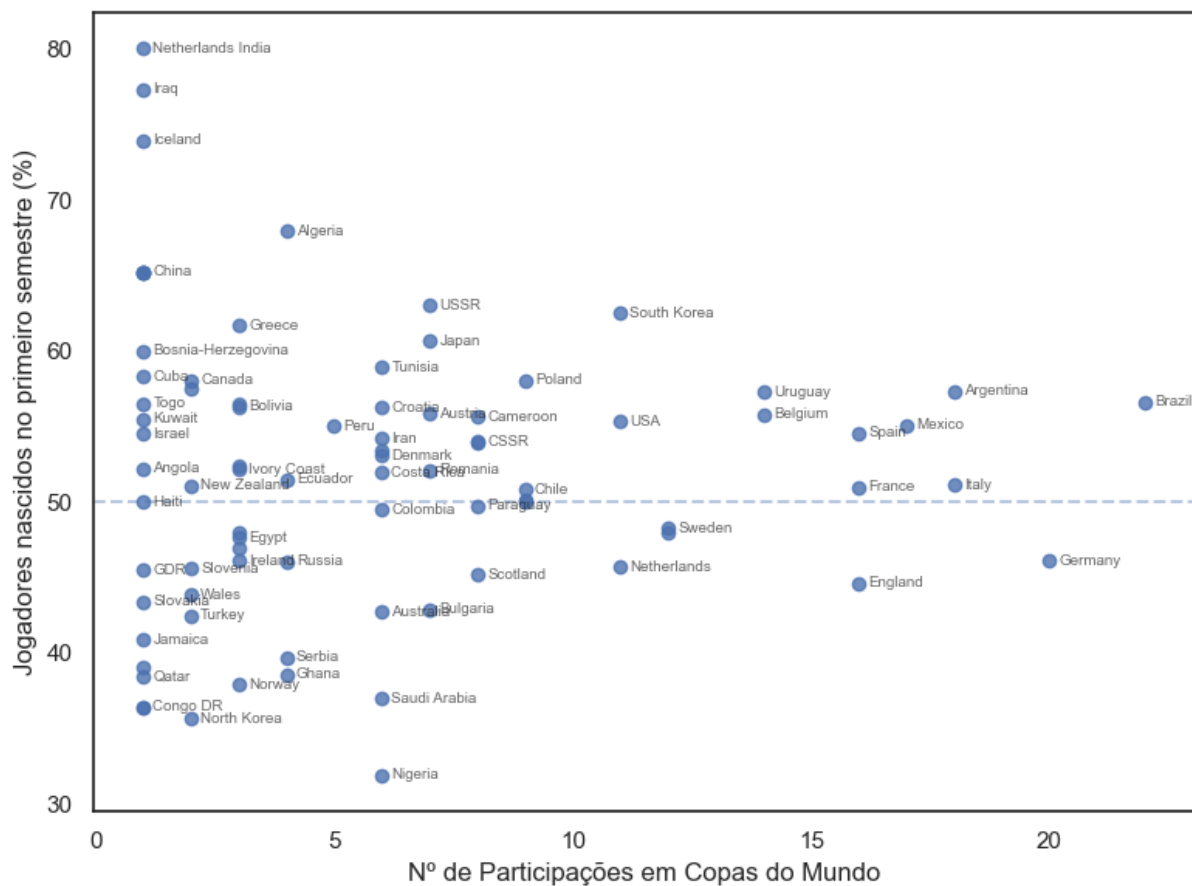
As features utilizadas foram: Idade, Mês de nascimento e Posição e os modelos de classificação foram treinados com Regressão Logística, RandomForest e XGBoost, garantindo que fosse possível realizar uma comparação desde o modelo mais simples até o mais complexo neste trabalho.

4 RESULTADOS

4.1 Análise Exploratória

Durante a fase de análise exploratória da pesquisa, um dos principais resultados foi observado ao plotar um scatter plot das seleções ao longo das copas do mundo. Comparou-se a quantidade de jogadores que nasceram no primeiro semestre de cada time com a quantidade de edições que cada seleção participou das copas. Foi verificado que os times com mais frequência nesses campeonatos também dispõem de uma quantidade de jogadores nascidos predominantemente no primeiro semestre, conforme a Figura 4 e a Tabela 3.

Figura 4 – Percentual de Jogadores nascido no primeiro semestre por time vs N° de Partidas



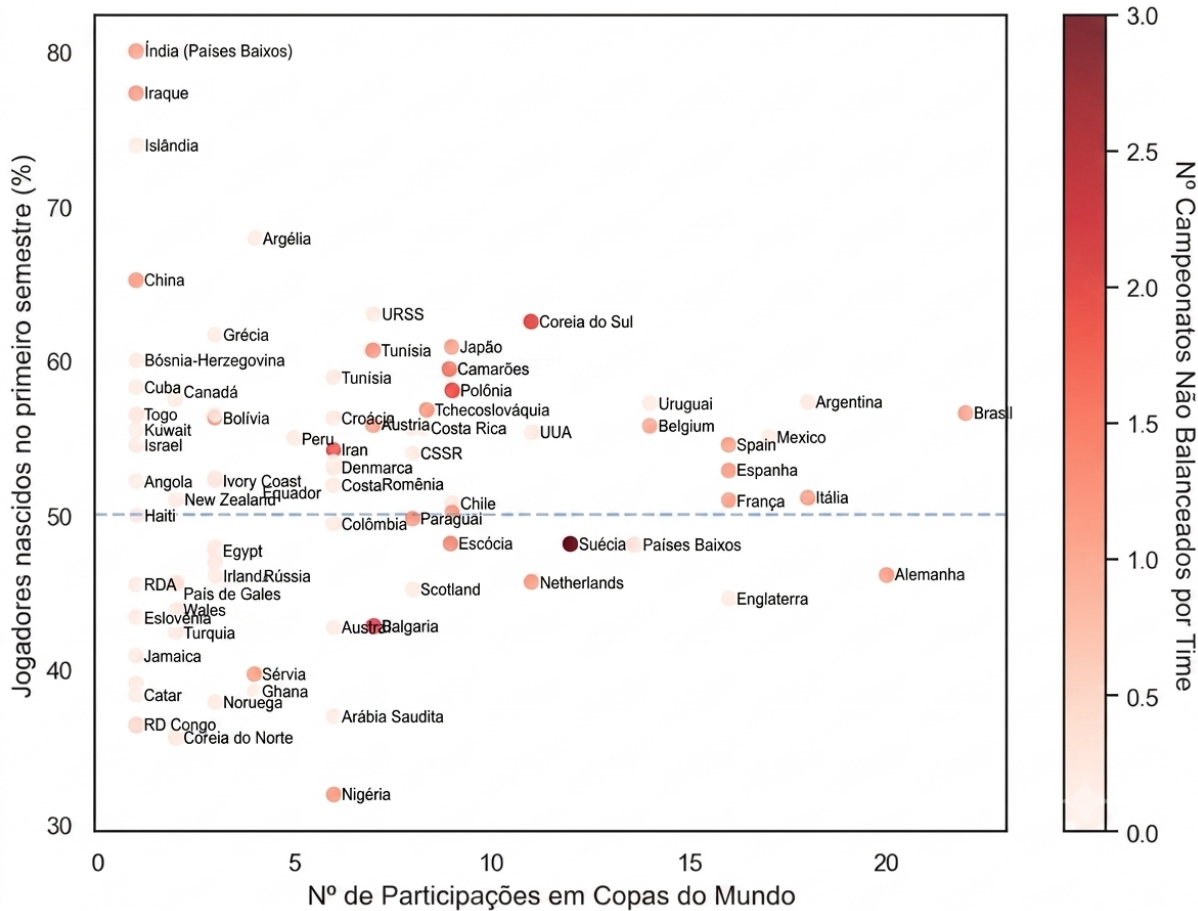
Fonte : (Autor, 2026)

Tabela 3 – Dados agregados por seleção para análise do EIR

Time	Edições	Média_Q1+Q2 (%)	Média_Q3+Q4 (%)
Brasil	22	56,59	43,40
Alemanha	20	46,11	53,91
Argentina	18	57,31	42,67
Itália	18	51,12	48,84
México	17	55,05	44,92
Inglaterra	16	44,61	55,36

Para validar estatisticamente se há de fato um desbalanceamento de classes que justifique o Efeito Idade Relativa, foram utilizados dois testes já consolidados na literatura do EIR: o teste do qui-quadrado e o teste de odds-ratio. Em ambos os testes quanto mais desbalanceada a classe maior é o gradiente da cor no gráfico que foi utilizado. Na Figura 5, não foi possível verificar uma tendência clara que relacione o desbalanceamento das seleções e a frequência nas Copas do Mundo. Seleções como o Japão que participou apenas de sete edições apresentaram em média um Efeito Idade Relativa maior que times como o Brasil, Argentina e Alemanha, que participaram de mais de 15 campeonatos. Vale ressaltar uma limitação do teste qui-quadrado que foi mencionada no referencial teórico: Um requisito mínimo para realização do teste do qui-quadrado é que cada time tenha pelo menos 20 jogadores ($\mu_{ij} \geq 0.25n$). Tendo cada seleção apenas 26 jogadores, o teste fica bem próximo do mínimo esperado.

Figura 5 – Teste do Qui-quadrado



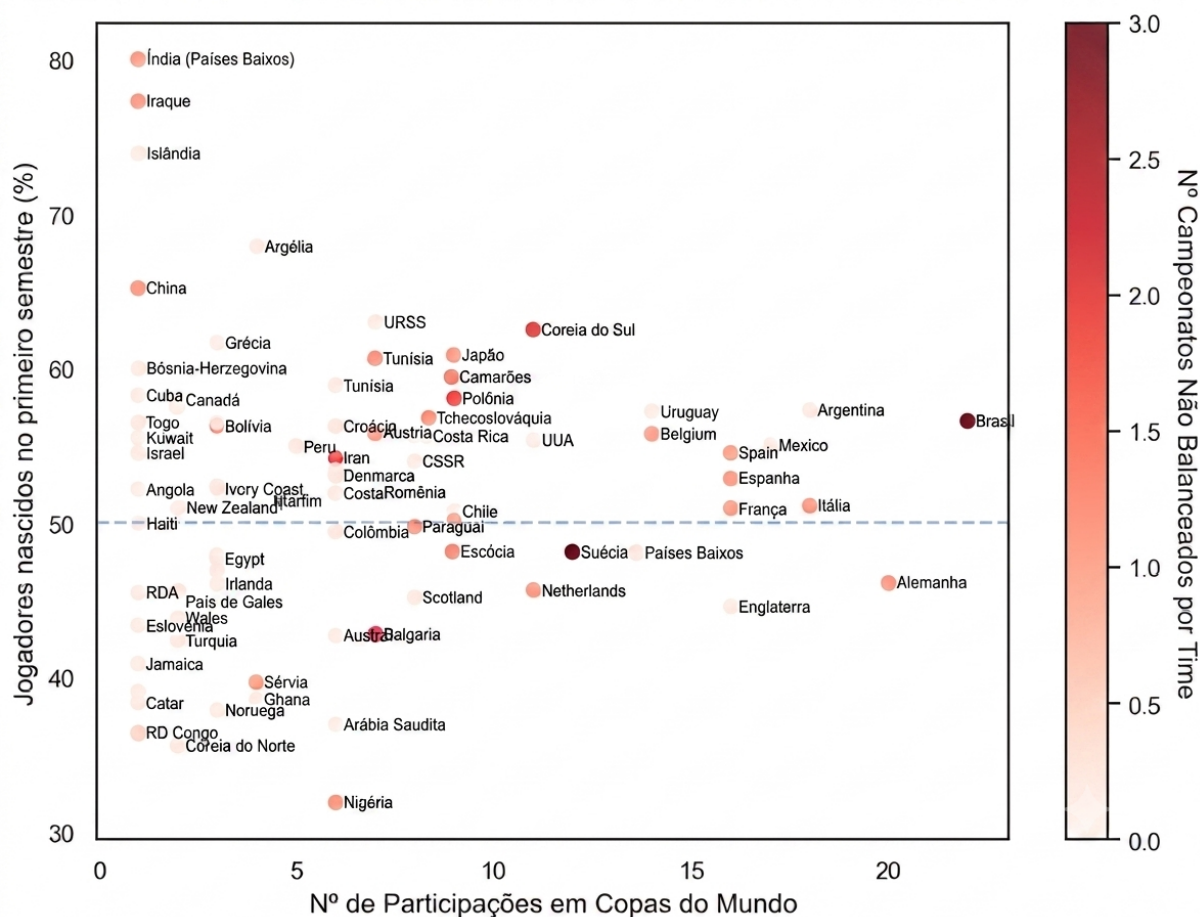
Por outro lado, o teste de odds ratio favoreceu o argumento do Efeito Idade Relativa. Na Tabela 4 pode-se verificar que dos dez times que mais jogaram nas edições da Copa do Mundo, seis deles tiveram mais de 50% dos campeonatos favorecendo jogadores nascidos no primeiro semestre do ano. No Brasil, por exemplo, dos 22 jogos em Copa do Mundo, 17 jogos tem um forte EIR favorecendo jogadores que nasceram no primeiro semestre do ano. Esses resultados podem ser também analisados graficamente na Figura 6.

Tabela 4 – Resultados de desequilíbrio por equipe (chi-square vs. odds_ratio)

Time	Edições	Qui-quadrado	Odds ratio
Brasil	22	1	17
Alemanha	20	1	6
Argentina	18	0	12
Itália	18	1	8
Mexico	17	0	11
Inglaterra	16	0	4
França	16	1	6
Espanha	16	1	9
Uruguai	14	0	11
Bélgica	14	1	9

Fonte : (Autor, 2026)

Figura 6 – Teste de Odds-ratio



Fonte : (Autor, 2026)

Adicionalmente, outro resultado interessante da Análise exploratória é encontrado ao estudar os jogadores, ao invés dos times, que mais participaram das edições da copa. Filtrou-se os jogadores que jogaram pelo menos quatro Copas do Mundo para analisar

quais posições e meses de nascimento são mais presentes e portanto podem influenciar na permanência dos mesmos nas próximas edições. O resultado obtido está ilustrado na Tabela 5 e na Tabela 6, em que pode-se verificar que os jogadores com maior frequência nasceram predominantemente no primeiro trimestre do ano conforme a hipótese do EIR.

Tabela 5 – Distribuição percentual de jogadores por trimestre

Trimestre	% Jogadores
Q1 (Jan-Mar)	30,2
Q2 (Abr-Jun)	23,3
Q3 (Jul-Set)	26,7
Q4 (Out-Dez)	19,8

Fonte : (Autor, 2026)

Adicionalmente, foi identificado que os jogadores mais frequentes na Copa do Mundo tem um EIR maior nas posições de Zagueiro e Goleiro em campo, o que corrobora a literatura(COTO-LOUSAS et al., 2025), visto que o EIR é predominante em posições que exigem mais fisicamente dos atletas por estar diretamente relacionado com a diferença de maturidade biológica presente entre os jogadores desde o início de suas carreiras (MALINA et al., 2000).

Tabela 6 – Distribuição percentual por posição e trimestre

Trimestre	Zagueiro	Atacante	Goleiro	Meio-campo
Q1 (Jan-Mar)	38,5	19,2	23,1	19,2
Q2 (Abr-Jun)	35,0	10,0	30,0	25,0
Q3 (Jul-Set)	26,1	21,7	30,4	21,7
Q4 (Out-Dez)	11,8	11,8	35,3	41,2

Fonte : (Autor, 2026)

4.2 Aprendizado de Máquina e Predição

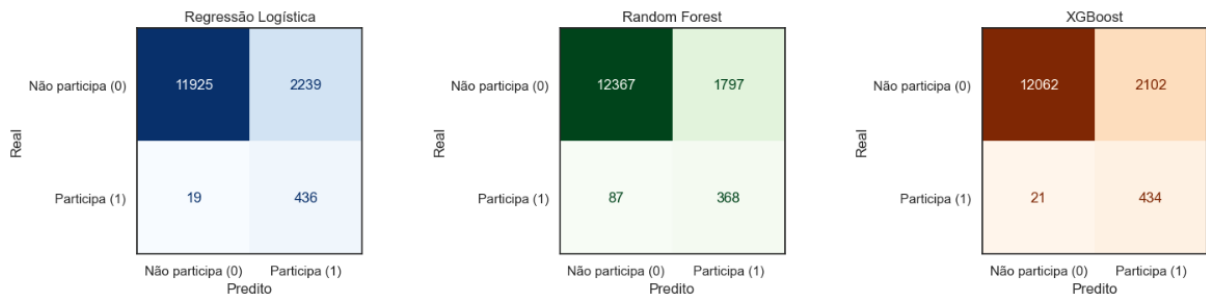
Durante o trabalho, treinaram-se três classificadores com o objetivo de estudar tendências do EIR nas próximas edições da Copa do Mundo. Foi verificado que a precisão dos modelos com relação a classe "Participa" dos jogadores para a Copa do Mundo de 2026 é bem menor que da classe "Não participa" conforme observado na Tabela 7. Parte desse resultado se deve ao desbalanceamento de classes. A quantidade de edições que cada jogador participou pode atingir no máximo cinco copas no melhor dos casos com jogadores como Lionel Messi ou Cristiano Ronaldo. Por outro lado, a quantidade de edições em que cada jogador não participou pode chegar a 21 no pior dos casos. Esse desbalanceamento resultou em um baixo F1-Score entre 0,2786 e 0,2902 que permaneceu apesar da aplicação de técnicas como a de *undersample* usando SMOTE.

Tabela 7 – Métricas de Performance dos Modelos

Modelo	Classe	Precisão	Recall	F1-Score	AUC
Regressão Logística	Não Participa	0.9984	0.8419	0.9114	0.9419
	Participa	0.1630	0.9582	0.2742	
	ROC-AUC	—	—	—	
Random Forest	Não Participa	0.9930	0.8731	0.9292	0.9002
	Participa	0.1700	0.8088	0.2809	
	ROC-AUC	—	—	—	
XGBoost	Não Participa	0.9983	0.8516	0.9166	0.9436
	Participa	0.1711	0.9538	0.2850	
	ROC-AUC	—	—	—	

Verificando a matriz de confusão da Figura 7, percebe-se que, de fato, os três modelos conseguem ter um bom desempenho na predição dos jogadores que não irão participar da Copa do Mundo de 2026 (Verdadeiro Negativo), ainda assim deixam a desejar com relação aos que irão participar (Verdadeiro Positivo).

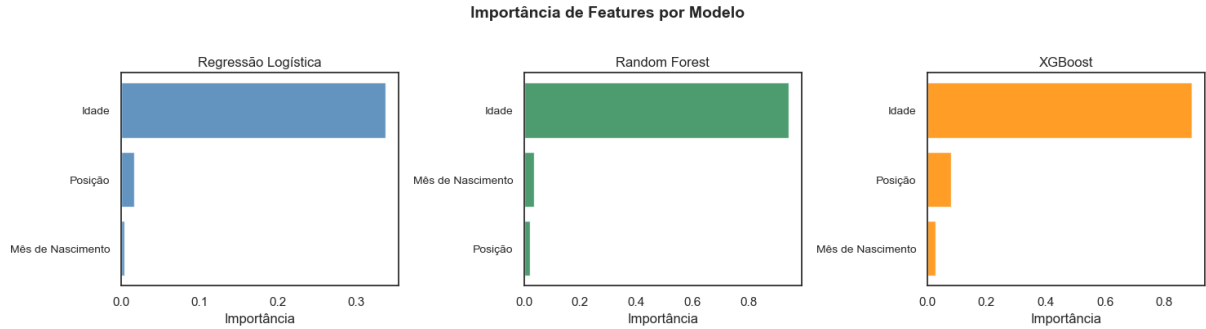
Figura 7 – Matriz de confusão



Fonte : (Autor, 2026)

Com a importância de *features* que é disponibilizado pela biblioteca *scipy* para cada um dos modelos, é possível identificar qual *feature* foi mais relevante para o treino. Pode-se observar que diferente do Random Forest e do XGBoost, o algoritmo de Regressão Logística utilizou principalmente o atributo de número de copas que cada jogador participou para predição. Para o contexto do EIR, apenas o classificador que utiliza Random Forest teve o mês de nascimento com uma certa relevância para predição.

Figura 8 – Importância de Features



Fonte : (Autor, 2026)

Utilizando os resultados de predição do modelo XGBoost, realizou-se novamente o teste do qui-quadrado e odds ratio. Na Tabela 8, verificou-se as dez seleções com o maior desbalanceamento. Pode-se perceber que alguns times como Polônia e Bélgica continuam com uma quantidade acentuada de jogadores no primeiro semestre e com um qui-quadrado acentuado, o que se alinha com a tendência histórica dessas seleções que foi apresentada na Figura 5. Outra observação é que majoritariamente os times que aparecem no rank possuem mais jogadores nascidos no primeiro semestre que no segundo. Um contraste está em equipes com Estados Unidos e Tunisia, que não apresentavam um forte EIR nas copas passadas e aparece no rank para 2026, mas com um p-valor alto, o que reduz significativamente o poder estatístico para essa seleção. Em conclusão, apenas a Polônia apresentou distribuição de nascimentos por trimestre estatisticamente significativa ($p < 0.05$).

Tabela 8 – Efeito EIR na edição de 2026 ordenado por qui-quadrado

Time	Q1+Q2	Q3+Q4	qui_quadrado	valor_p
Polônia	20	6	8.769231	0.033
Coreia do Sul	10	16	7.538462	0.057
Tunísia	16	10	6.615385	0.085
Canadá	16	10	5.384615	0.146
França	16	9	5.240000	0.155
Bélgica	18	8	5.076923	0.166
Senegal	16	10	4.461538	0.216
EUA	10	16	4.461538	0.216
Espanha	15	11	4.461538	0.216
Irã	16	9	4.280000	0.233

Para odds-ratio, a Polônia continua sendo a dominante com o maior odds-ratio, ainda que não tão representada na Figura 6. Seleções como Argentina e Uruguai seguiram a tendência apresentada nas edições anteriores da copa. Uma observação é que países como

Brasil, México e Itália que continham um forte EIR histórico pela métrica de odds-ratio não são representados no rank pelo modelo.

Tabela 9 – EIR no campeonato de 2026 ordenada por odds-ratio

TIME	Q1+Q2	Q3+Q4	odds ratio
Polônia	20	6	3.333333
Argentina	18	8	2.250000
Bélgica	18	8	2.250000
Costa Rica	17	9	1.888889
Holanda	17	9	1.888889
Uruguai	17	9	1.888889
Irã	16	9	1.777778
França	16	9	1.777778
Senegal	16	10	1.600000
Alemanha	16	10	1.600000

O XGBoost obteve ROC-AUC de 0,941 (o melhor entre os três modelos) e recall de 0,928 para a classe "Participou", demonstrando boa capacidade de identificar jogadores com histórico de convocação continuada. O mês de nascimento compõe o vetor de features e, embora com importância relativa menor que idade e posição, sua influência acumulada ao longo de múltiplas copas históricas está refletida nas predições: o modelo aprendeu, a partir dos dados reais, que jogadores nascidos no primeiro semestre têm ligeiramente maior probabilidade de permanecer nas seleções reproduzindo, nas predições, o EIR observado nos dados históricos de treinamento.

5 CONCLUSÃO

Este projeto de pesquisa propôs a investigação do Efeito da Idade Relativa (EIR) no contexto do futebol profissional, com ênfase nas Copas do Mundo, utilizando técnicas estatísticas e métodos de aprendizado de máquina. A partir da revisão da literatura, foi possível constatar que o EIR é um fenômeno amplamente documentado nas categorias de base e que seus efeitos podem se estender ao futebol profissional por meio de um efeito cascata, embora ainda existam lacunas na literatura quando se considera uma base de dados ampla e longitudinal, como a das Copas do Mundo.

A metodologia proposta foi estruturada de forma a permitir uma análise abrangente do fenômeno, iniciando-se pela fundamentação teórica e análise exploratória dos dados, passando pelo tratamento das informações, até a aplicação e avaliação de modelos preditivos. A análise exploratória preliminar já indicou padrões relevantes, sugerindo uma predominância de jogadores nascidos no primeiro semestre em seleções com maior histórico de participações no torneio, o que reforça a hipótese da presença do EIR nesse contexto.

Entre as limitações do trabalho, pôde-se perceber um desafio considerável no tratamento com dados desbalanceados dentro da temática do EIR. Ainda com o uso de técnicas específicas para esses casos como SMOTE, não foi possível obter uma precisão significativa para os três modelos utilizados, que seguem desde o modelo mais tradicional a um dos modelos mais complexos da literatura. Dessa forma, sugere-se que em novos trabalhos, seja utilizado diferentes métodos para tratar com esse desbalanceamento tal como técnicas de ajuste de peso das classes majoritárias e minoritárias discutidas.

Por fim, espera-se que os resultados obtidos ao longo da continuidade deste trabalho contribuam tanto para o avanço acadêmico na área de análise de dados aplicada ao esporte quanto para a discussão prática sobre critérios de seleção e desenvolvimento de atletas no futebol profissional. Ao aprofundar o entendimento sobre o Efeito da Idade Relativa nas Copas do Mundo, este estudo busca fornecer subsídios para reflexões futuras sobre políticas mais justas e eficientes no processo de formação e seleção de jogadores.

Referências

- AGRESTI, A. *An Introduction to Categorical Data Analysis*. 2nd. ed. Hoboken, NJ: John Wiley & Sons, 2007. ISBN 978-0-471-22618-5. Citado 3 vezes nas páginas 14, 15 e 21.
- ALBUQUERQUE, M. R.; FRANCHINI, E.; LAGE, G. M.; Da Costa, V. T.; COSTA, I. T.; MALLOY-DINIZ, L. F. The relative age effect in combat sports: An analysis of olympic judo athletes, 1964-2012. *Perceptual and Motor Skills*, v. 121, n. 1, p. 300–308, 2015. ISSN 1558688X. Citado na página 8.
- ALTMANN, S.; RUF, L.; THIEM, S.; BECKMANN, T.; WOHAK, O.; ROMEIKE, C.; HÄRTEL, S. Prediction of talent selection in elite male youth soccer across 7 seasons: A machine-learning approach. *Journal of Sports Sciences*, v. 42, n. 24, p. 2481–2494, 2024. Citado na página 11.
- BARNESLEY, R. Hockey success and birthdate: The relative age effect Exploring the Relative Age Effect View project. n. January, p. 23–28, 1985. Disponível em: <<https://www.researchgate.net/publication/284328248>>. Citado 2 vezes nas páginas 8 e 12.
- BUSSAB, W. d. O.; MORETTIN, P. A. *Estatística Básica*. 6. ed. São Paulo: Editora Saraiva, 2009. Citado na página 14.
- CARLI, G. C.; LUGUETTI, C. N.; RÉ, A. H. N.; BÖHME, M. T. S. *Artigo Original Efeito da idade relativa no futebol Relative age effect in soccer players*. [S.l.], 2009. v. 17, n. 3, 25–31 p. Disponível em: <<http://www.fifa.com>>. Citado na página 13.
- CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2016. p. 785–794. Citado na página 20.
- COTO-LOUSAS, J.; SIERRA-DÍAZ, J.; SAIZ-GONZÁLEZ, P.; GONZÁLEZ-VÍLLORA, S. Timing is everything: An analysis of the relative age effect in a professional football club. *Football Studies*, 2025. Citado na página 29.
- DELORME, N. Do weight categories prevent athletes from relative age effect? *Journal of Sports Sciences*, v. 32, n. 1, p. 16–21, 2014. ISSN 02640414. Citado na página 8.
- DUARTE, D. N. *Desenvolvimento de Web Scraping e Dashboard para Análise do Efeito da Idade Relativa no Futebol Brasileiro*. 48 p. Trabalho de Conclusão de Curso (Graduação) — Universidade Federal do Maranhão, São Luís, 2025. Citado na página 9.
- FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001. Citado na página 20.
- GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2. ed. Sebastopol, CA: O'Reilly Media, 2019. Citado 4 vezes nas páginas 18, 19, 20 e 21.
- HEIM:SPIEL. *WorldFootball*. 2000–2023. Disponível em: <<https://www.worldfootball.net/>>. Acesso em: 27 mar. 2023. Citado na página 23.

HELSEN, W. F.; STARKES, J. L.; Van Winckel, J. The Influence of Relative Age on Success and Dropout in Male Soccer Players. *American Journal of Human Biology*, v. 10, n. 6, p. 791–798, 1998. ISSN 10420533. Citado 3 vezes nas páginas 8, 12 e 13.

HO, T. K. Random decision forests. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR)*. [S.l.]: IEEE, 1995. v. 1, p. 278–282. Citado na página 19.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. 3rd. ed. Hoboken, NJ: John Wiley & Sons, 2013. (Wiley Series in Probability and Statistics). Citado 2 vezes nas páginas 16 e 17.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An Introduction to Statistical Learning: with Applications in R*. 2nd. ed. New York: Springer, 2021. ISBN 978-1-0716-1417-4. Citado na página 19.

JAUHIAINEN, S.; ÄYRÄMÖ, S.; FORSMAN, H.; KAUPPI, J.-P. Talent identification in soccer using a one-class support vector machine. *International Journal of Computer Science in Sport*, v. 18, n. 3, p. 125–136, 2019. Citado 2 vezes nas páginas 10 e 11.

KELLY, A. L.; WILLIAMS, C. A.; COOK, R.; SÁIZ, S. L. J.; WILSON, M. R. A multidisciplinary investigation into the talent development processes at an English football academy: A machine learning approach. *Sports*, v. 10, n. 10, p. 159, 2022. Citado 2 vezes nas páginas 10 e 11.

MALINA, R. M.; PEÑAREYES, M. E.; EISENMANN, J. C.; HORTA, L.; RODRIGUES, J.; MILLER, R. Height, mass and skeletal maturity of elite portuguese soccer players aged 11-16 years. *Journal of Sports Sciences*, v. 18, n. 9, p. 685–693, 2000. ISSN 1466447X. Citado 3 vezes nas páginas 8, 13 e 29.

MUJICA, I.; VAEYENS, R.; MATTHYS, S. P.; SANTISTEBAN, J.; GOIRIENA, J.; PHILIPPAERTS, R. The relative age effect in a professional football club setting. *Journal of Sports Sciences*, v. 27, n. 11, p. 1153–1158, 2009. ISSN 02640414. Citado 2 vezes nas páginas 8 e 13.

NIELSEN, A. *Practical Time Series Analysis: Previsão com Estatística e Machine Learning*. 1. ed. Rio de Janeiro: Alta Books, 2021. ISBN 9786555201895. Citado na página 22.

PADRÓN-CABO, A.; REY, E.; GARCÍA-SOIDÁN, J. L.; PENEDO, E.; PADRÓN-CABO, A.; REY, E.; GARCÍA-SOIDÁN, J. L.; PENEDO, E. Large Scale Analysis of Relative Age Effect on Professional Soccer Players in FIFA Designated Zones Large Scale Analysis of Relative Age Effect on Professional Soccer Players in FIFA Designated Zones . v. 8668, n. August, 2017. Citado 2 vezes nas páginas 9 e 14.

REIS, *Desenvolvimento de Webscrapping e Dashboard para análise do Efeito Idade Relativa*. Trabalho de Conclusão de Curso (Graduação) — Universidade Federal do Maranhão, São Luís, 2023. Citado 2 vezes nas páginas 9 e 22.

SAAVEDRA-GARCÍA, M.; GUTIÉRREZ-AGUILAR, Ó.; SA-MARQUES, P.; FERNÁNDEZ-ROMERO, J. J. Efecto de la edad relativa en el atletismo Español. *Cuadernos de Psicología del Deporte*, v. 16, n. 1, p. 275–286, 2016. ISSN 15788423. Citado na página 9.

SAAVEDRA-GARCÍA, M.; MATABUENA, M.; MONTERO-SEOANE, A.; FERNÁNDEZ-ROMERO, J. J. Relative age effect in FIFA football tournaments (1908–2012): A new approach with additive logistic regression models. *PLOS ONE*, v. 14, 2019. Citado na página 10.

SALINERO, J. J.; PÉREZ, B.; BURILLO, P.; LESMA, M. L. Relative age effect in European professional football. Analysis by position. *Journal of Human Sport and Exercise*, v. 8, n. 4, p. 966–973, 2013. ISSN 19885202. Citado 2 vezes nas páginas 8 e 9.

SIMMONS, C.; PAULL, G. C. Season-of-birth bias in association football. *Journal of Sports Sciences*, v. 19, n. 9, p. 677–686, 2001. ISSN 02640414. Citado 3 vezes nas páginas 8, 12 e 13.

VITORIA, J. *Efeito Idade Relativa nas Copas do Mundo da FIFA : Uma Análise Exploratória com Dashboard Interativo*. 39 p. Trabalho de Conclusão de Curso (Graduação) — Universidade Federal do Maranhão, São Luís, 2025. Citado na página 9.

YAGÜE, J. M.; RUBIA, A. de la; SÁNCHEZ-MOLINA, J.; MAROTO-IZQUIERDO, S.; MOLINERO, O. The relative age effect in the 10 best leagues of male professional football of the union of european football associations (UEFA). *Journal of Sports Science and Medicine*, v. 17, n. 3, p. 409–416, 2018. ISSN 13032968. Citado na página 9.