UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

EVZEN PHARDIN BUSTINZA CALLE

Modelagem interpretável do risco de lesão em atletas com aprendizado de máquina e análise de fatores de risco

Campinas
24/06/2026

EVZEN PHARDIN BUSTINZA CALLE

Modelagem interpretável do risco de lesão em atletas com aprendizado de máquina e análise de fatores de risco

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do(a) Prof. Laércio Luís Vendite.

Resumo

Este trabalho desenvolve uma abordagem interpretável de aprendizado de máquina para estimar a ocorrência de lesão na temporada seguinte em jogadores universitários de futebol. Foi utilizado o *University Football Injury Prediction Dataset*, composto por 800 jogadores e 18 variáveis relacionadas a características físicas, esportivas, funcionais e comportamentais. Após a auditoria e a análise exploratória dos dados, a base foi dividida de forma estratificada em conjuntos de treinamento e teste. Foram comparados um classificador de referência, regressão logística regularizada, máquina de vetores de suporte e Random Forest por meio de validação cruzada aninhada. A regressão logística foi selecionada por apresentar desempenho comparável aos modelos não lineares e maior transparência. No conjunto de teste, o modelo obteve ROC-AUC de 0,9978, sensibilidade de 0,9750, especificidade de 0,9875 e acurácia balanceada de 0,9813. Estresse, sono, equilíbrio, tempo de reação e desempenho no sprint foram as características de maior influência. Apesar do elevado desempenho, os resultados representam uma validação interna em uma única base pública e não demonstram relações causais ou aplicabilidade clínica imediata. Conclui-se que modelos multivariados interpretáveis podem auxiliar na identificação de padrões associados às lesões, desde que utilizados de forma complementar à avaliação dos profissionais do esporte e da saúde.

Palavras-chave: lesões esportivas; futebol; aprendizado de máquina; regressão logística; interpretabilidade.

Abstract

This study develops an interpretable machine learning approach to estimate injury occurrence in the following season among university football players. The University Football Injury Prediction Dataset, composed of 800 players and 18 variables related to physical, sporting, functional, and behavioral characteristics, was used. After data auditing and exploratory analysis, the dataset was stratified into development and test sets. A baseline classifier, regularized logistic regression, support vector machine, and Random Forest were compared using nested cross-validation. Logistic regression was selected because it achieved performance comparable to nonlinear models while providing greater transparency. On the test set, the model obtained a ROC-AUC of 0.9978, sensitivity of 0.9750, specificity of 0.9875, and balanced accuracy of 0.9813. Stress, sleep duration, balance, reaction time, and sprint performance were the most influential characteristics. Despite the high performance, the results represent internal validation on a single public dataset and do not establish causal relationships or immediate clinical applicability. Interpretable multivariate models may help identify injury-associated patterns, provided that they complement the assessment of sports and healthcare professionals.

Keywords: sports injuries; football; machine learning; logistic regression; interpretability.

Conteúdo

1 Introdução

No futebol, lesões afetam diretamente a disponibilidade dos atletas e o planejamento das equipes. Além das consequências individuais, períodos de afastamento podem alterar a continuidade dos treinamentos, a composição do elenco em partidas e a tomada de decisão da comissão técnica. Estudos com jogadores profissionais mostram que lesões são eventos frequentes ao longo da temporada e podem gerar períodos relevantes de indisponibilidade [3].

A previsão desse tipo de evento é difícil porque o risco de lesão não depende de uma única característica do jogador. Modelos de etiologia descrevem a lesão esportiva como resultado de um processo dinâmico, no qual fatores individuais, histórico clínico, capacidades físicas, exposição ao esporte, recuperação e eventos da prática interagem ao longo do tempo [9]. Por isso, variáveis analisadas isoladamente tendem a representar apenas parte do problema.

Métodos de aprendizado de máquina são úteis nesse contexto por permitirem a análise conjunta de variáveis físicas, funcionais, esportivas e comportamentais. Ainda assim, revisões sobre previsão de lesões no futebol ressaltam que o desempenho desses modelos depende da população estudada, da definição de lesão, das variáveis disponíveis e do procedimento de validação empregado [8]. Assim, alto desempenho preditivo não deve ser interpretado automaticamente como evidência de aplicabilidade prática.

Além da capacidade de previsão, a interpretabilidade é importante em aplicações relacionadas à saúde e ao desempenho esportivo. Um modelo que apenas produz uma probabilidade, sem indicar quais características sustentam essa estimativa, tem utilidade limitada para apoiar discussões entre analistas, preparadores físicos, médicos e fisioterapeutas.

Este trabalho utiliza o *University Football Injury Prediction Dataset*, disponibilizado na plataforma Kaggle e composto por informações de 800 jogadores universitários de futebol [13]. A base contém variáveis relacionadas a características físicas, informações esportivas, testes funcionais, estilo de vida e adesão à rotina de aquecimento. O desfecho analisado indica se o jogador apresentou ou não lesão na temporada seguinte.

O objetivo deste trabalho é desenvolver e avaliar uma abordagem interpretável

de aprendizado de máquina para estimar a ocorrência de lesão na temporada seguinte. Os resultados são tratados como padrões preditivos observados na base analisada, sem a pretensão de estabelecer relações causais, fornecer diagnóstico individual ou substituir a avaliação de profissionais do esporte e da saúde.

1.1 Problema de pesquisa

A questão central do trabalho é:

> Em que medida características físicas, esportivas, funcionais e comportamentais podem ser utilizadas por modelos de aprendizado de máquina para estimar, de maneira interpretável, a ocorrência de lesão na temporada seguinte em jogadores universitários de futebol?

1.2 Objetivos

O objetivo geral consiste em desenvolver e avaliar uma abordagem interpretável de aprendizado de máquina para estimar a ocorrência de lesão na temporada seguinte.

Os objetivos específicos são:

1. descrever e analisar as características da base de dados;
2. comparar o desempenho da regressão logística com modelos capazes de representar relações não lineares;
3. avaliar a discriminação, a classificação e a qualidade das probabilidades produzidas pelos modelos;
4. identificar as variáveis de maior influência nas estimativas;
5. discutir as limitações e as possibilidades de aplicação dos resultados.

Os resultados são interpretados como associações observadas na base de dados. O modelo não tem como finalidade estabelecer relações causais, fornecer diagnóstico ou substituir a avaliação de profissionais da saúde e do esporte.

2 Contextualização

2.1 Lesões esportivas e fatores associados

A ocorrência de lesões no futebol envolve fatores intrínsecos e extrínsecos. Entre os elementos frequentemente discutidos estão histórico de lesões, capacidades físicas, controle neuromuscular, exposição competitiva, carga de treinamento, sono, estresse e recuperação. Essa combinação reforça que o risco deve ser entendido de forma multifatorial.

O histórico de lesões é um dos fatores mais consistentes na literatura. Em estudo prospectivo com jogadores de futebol, atletas lesionados em uma temporada apresentaram maior risco de sofrer nova lesão na temporada seguinte [5].

Aspectos psicossociais e de recuperação também são relevantes. Uma meta-análise identificou associação entre respostas ao estresse, histórico de estressores e ocorrência de lesões esportivas [6]. O sono, por sua vez, participa de processos de recuperação física e cognitiva; em jogadores de futebol, sono insuficiente ou de baixa qualidade tem sido relacionado a pior desempenho e a maior ocorrência ou gravidade de lesões, embora a literatura ainda apresente resultados heterogêneos [2].

Programas estruturados de aquecimento e treinamento neuromuscular têm sido investigados como estratégias preventivas [4, 12]. Ainda assim, testes físicos isolados apresentam capacidade limitada para identificar, de forma individual, quais atletas sofrerão lesões futuras [11]. Essa limitação justifica o uso de abordagens multivariadas e interpretações cautelosas.

2.2 Aprendizado de máquina na previsão de lesões

O aprendizado de máquina supervisionado utiliza observações com desfechos conhecidos para construir uma função capaz de produzir estimativas para novas observações. Neste trabalho, o problema é formulado como uma tarefa de classificação binária.

Considere o conjunto de dados

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \tag{1}$$

em que $\mathbf{x}_i$ representa as características do jogador i e

$$Y_i = \begin{cases} 0, & \text{se o jogador não apresentou lesão;} \\ 1, & \text{se o jogador apresentou lesão.} \end{cases} \tag{2}$$

O modelo busca estimar

$$\widehat{p}_i \approx P\left(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i\right). \tag{3}$$

A probabilidade pode ser transformada em uma classificação por meio de um limiar τ:

$$\widehat{y}_i = \begin{cases} 1, & \text{se } \widehat{p}_i \geq \tau; \\ 0, & \text{se } \widehat{p}_i < \tau. \end{cases} \tag{4}$$

Na previsão de lesões no futebol, modelos de aprendizado de máquina têm sido utilizados para integrar variáveis de diferentes naturezas. A comparação entre estudos, porém, é dificultada por diferenças nas populações, nas definições de lesão, nas variáveis disponíveis e nos procedimentos de validação [8].

2.3 Interpretabilidade

A interpretabilidade corresponde à possibilidade de compreender como as características dos jogadores influenciam as estimativas produzidas pelo modelo.

Na regressão logística, a probabilidade estimada é dada por

$$\widehat{p}_i = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)\right]}. \tag{5}$$

Os coeficientes podem ser convertidos em razões de chances:

$$\mathrm{OR}_j = \exp(\beta_j). \tag{6}$$

Valores de OR > 1 indicam associação com maiores chances estimadas de lesão, enquanto valores de OR < 1 indicam associação com menores chances, mantendo

constantes as demais variáveis.

Modelos mais complexos podem exigir métodos adicionais de explicação, como valores SHAP. No presente trabalho, a regressão logística foi considerada especialmente relevante por combinar desempenho preditivo e interpretação direta.

3 Dados e preparação

3.1 Tratamento e auditoria dos dados

A base contém 800 observações e 19 colunas. Cada observação corresponde a um jogador universitário de futebol. Dezoito colunas são utilizadas como variáveis preditoras e uma corresponde ao desfecho `Injury_Next_Season`.

As variáveis foram organizadas nos seguintes grupos:

- **características físicas:** idade, altura, massa corporal e índice de massa corporal;
- **informações esportivas:** posição, horas semanais de treinamento, partidas disputadas e histórico de lesões;
- **aptidão física:** força do joelho, flexibilidade, tempo de reação, equilíbrio, velocidade no sprint e agilidade;
- **estilo de vida:** sono, estresse e qualidade da alimentação;
- **adesão preventiva:** realização da rotina de aquecimento.

A Tabela 1 apresenta o resumo da auditoria.

Tabela 1: Resumo da auditoria da base de dados.

Propriedade	**Resultado**
Número de observações	800
Número total de colunas	19
Variáveis preditoras	18
Variáveis categóricas	1
Valores ausentes	0
Valores infinitos	0
Registros duplicados	0
Variáveis constantes	0
Jogadores sem lesão	400
Jogadores com lesão	400

Não foram identificados valores ausentes, infinitos, registros duplicados ou variáveis constantes. Consequentemente, não foi necessário remover observações nem aplicar técnicas de imputação.

A variável `Position` foi tratada como categórica nominal. Já as variáveis de desfecho e de adesão ao aquecimento, respectivamente `Injury_Next_Season` e `Warmup_Routine_Adherence`, foram tratadas como indicadores binários.

Como o desfecho apresenta 400 observações em cada classe, não foram aplicadas técnicas artificiais de balanceamento. Essa distribuição perfeitamente simétrica, entretanto, é uma característica metodológica importante da base. A documentação pública do conjunto de dados o apresenta como uma base balanceada, mas não detalha se a proporção resulta de amostragem, seleção de casos e controles, balanceamento posterior ou construção sintética/curada dos dados. O estudo que utiliza essa mesma base também destaca a razão aproximadamente 1:1 entre jogadores lesionados e não lesionados como uma propriedade favorável à comparação de modelos, mas reconhece limitações de generalização por se tratar de uma única base pública [7].

Dessa forma, a frequência de 50% de lesões deve ser interpretada apenas como a composição do conjunto analisado, e não como uma estimativa da incidência real de lesões em jogadores universitários de futebol. Em estudos prospectivos de vigilância, a frequência observada dependeria da definição operacional de lesão, do tempo de acompanhamento, da exposição a treinos e partidas e das características da população. Portanto, o balanceamento facilita o treinamento e a comparação dos classificadores, mas reduz a

interpretação epidemiológica direta do desfecho e pode contribuir para um desempenho preditivo superior ao esperado em dados reais de clubes.

3.2 Análise exploratória dos dados

As distribuições das variáveis foram examinadas por meio de histogramas, boxplots, estatísticas descritivas e comparações entre os grupos com e sem lesão. Também foram calculados tamanhos de efeito, correlações ponto-bisseriais, testes qui-quadrado e fatores de inflação da variância.

A Figura 1 apresenta as diferenças padronizadas entre os grupos.

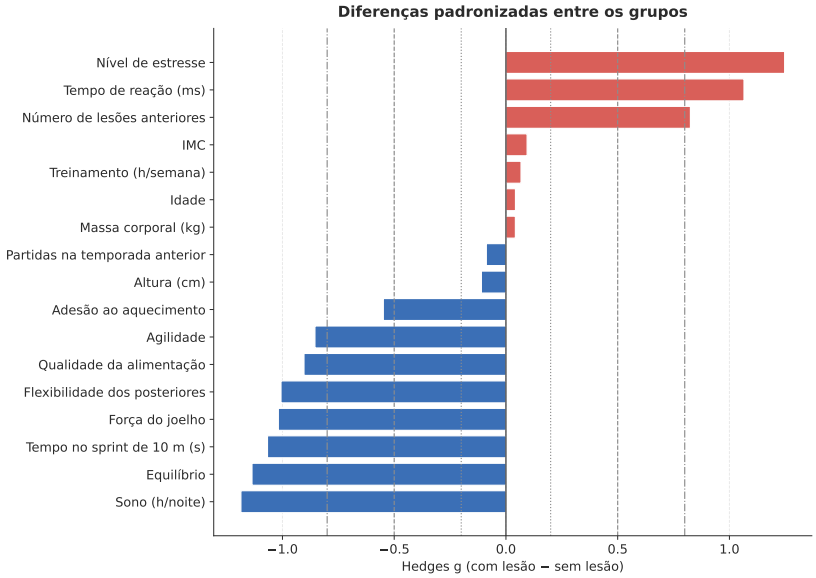


Figura 1: Diferenças padronizadas entre os jogadores com e sem lesão na temporada seguinte.

As maiores diferenças foram observadas para estresse, sono, equilíbrio, tempo de reação, velocidade no sprint, força do joelho, flexibilidade, alimentação, agilidade e histórico de lesões.

O nível de estresse apresentou correlação positiva com o desfecho, enquanto sono e equilíbrio apresentaram correlações negativas. O histórico de lesões também foi maior entre os jogadores classificados como lesionados.

A posição em campo apresentou associação de pequena magnitude com o desfecho:

$$\chi^2(3) = 2{,}422, \qquad p = 0{,}489, \qquad V = 0{,}055. \tag{7}$$

Em contrapartida, a adesão à rotina de aquecimento apresentou associação inversa com a ocorrência de lesão:

$$\chi^2(1) = 55{,}142, \qquad p < 0{,}001, \qquad V = 0{,}263. \tag{8}$$

A razão de chances não ajustada dos jogadores aderentes em comparação aos não aderentes foi

$$\mathrm{OR} = 0{,}329. \tag{9}$$

Esse resultado representa uma associação observada na base e não uma estimativa causal do efeito do aquecimento.

A análise de multicolinearidade identificou redundância severa entre altura, massa corporal e IMC. Foram obtidos os seguintes valores:

- IMC: VIF = 152,264;
- massa corporal: VIF = 114,933;
- altura: VIF = 41,155.

Para evitar instabilidade na interpretação da regressão logística, a especificação principal manteve altura e massa corporal e retirou o IMC. Uma especificação alternativa contendo o IMC foi posteriormente avaliada em uma análise de sensibilidade.

3.3 Separação e pré-processamento dos dados

A base foi dividida de forma estratificada em:

- conjunto de treinamento: 640 observações;
- conjunto de teste: 160 observações.

Foi utilizada semente aleatória igual a 42. Ambos os conjuntos mantiveram a proporção de 50% de jogadores em cada classe. Essa estratificação preserva a composição balanceada da base, mas não corrige a possível artificialidade dessa proporção; ela apenas

impede que o conjunto de treinamento e o conjunto de teste fiquem com prevalências distintas do desfecho.

Tabela 2: Separação entre os conjuntos de treinamento e teste.

Conjunto	**Total**	**Sem lesão**	**Com lesão**
Treinamento	640	320	320
Teste	160	80	80

As variáveis numéricas foram padronizadas pela transformação

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \tag{10}$$

em que μ_j e σ_j foram estimados exclusivamente nos dados de treinamento.

A posição foi transformada por codificação indicadora (*one-hot encoding*). O pré-processamento e os modelos foram integrados por meio de objetos `Pipeline`, de forma que todas as transformações fossem ajustadas novamente em cada etapa da validação cruzada.

O procedimento adotado pode ser resumido por

$$\text{dados} \longrightarrow \text{pré-processamento} \longrightarrow \text{modelo} \longrightarrow \text{probabilidade estimada.} \tag{11}$$

O conjunto de teste permaneceu isolado durante a comparação dos modelos, a seleção dos hiperparâmetros, a escolha do limiar e o ajuste das transformações.

4 Modelagem e resultados

4.1 Modelos aplicados

Foram comparados quatro classificadores: um modelo de referência, regressão logística regularizada, máquina de vetores de suporte com núcleo RBF e Random Forest. A escolha desses métodos permitiu comparar um modelo linear e diretamente interpretável com modelos capazes de representar relações não lineares.

4.1.1 Classificador de referência

O classificador de referência não utiliza as características dos jogadores para realizar as previsões. Sua classificação é baseada somente na distribuição observada das classes no conjunto de treinamento.

Esse modelo representa o nível mínimo de desempenho esperado. Dessa forma, um modelo de aprendizado de máquina somente é considerado útil se apresentar resultados superiores aos obtidos pelo classificador de referência.

Como a base analisada possui a mesma quantidade de jogadores com e sem lesão, o classificador de referência apresenta desempenho próximo ao de uma escolha aleatória, com ROC-AUC e acurácia balanceada iguais a aproximadamente 0,5.

4.1.2 Regressão logística regularizada

A regressão logística é um modelo de classificação que estima a probabilidade de ocorrência de um evento binário. Neste trabalho, ela estima a probabilidade de um jogador apresentar lesão na temporada seguinte.

Inicialmente, o modelo calcula uma combinação linear das características do jogador:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{12}$$

em que β_0 é o intercepto, β_j é o coeficiente associado à variável j e x_{ij} é o valor dessa variável para o jogador i.

A combinação linear é transformada em uma probabilidade por meio da função logística:

$$\widehat{p_i} = \frac{1}{1 + \exp(-\eta_i)}. \tag{13}$$

A classificação final é obtida comparando $\widehat{p_i}$ com um limiar de decisão τ.

Foi utilizada regularização L2, que adiciona uma penalização para coeficientes de grande magnitude. De forma simplificada, o ajuste busca minimizar

$$\mathcal{L}(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{14}$$

em que $\mathcal{L}$ representa a perda logística e λ controla a intensidade da penalização.

A regularização reduz a possibilidade de sobreajuste e contribui para a estabilidade dos coeficientes, especialmente quando existem variáveis correlacionadas. Na implementação utilizada, o parâmetro C é inversamente relacionado à penalização: valores menores de C representam regularização mais intensa.

Uma vantagem importante da regressão logística é a possibilidade de interpretar seus coeficientes por meio das razões de chances, conforme discutido na Seção 2.

4.1.3 Máquina de vetores de suporte com núcleo RBF

A máquina de vetores de suporte, ou SVM, busca construir uma fronteira de decisão que separe as classes com a maior margem possível. Os pontos mais próximos dessa fronteira são chamados de vetores de suporte e possuem maior influência na determinação do limite de classificação.

Quando os dados não podem ser adequadamente separados por uma fronteira linear, pode ser utilizada uma função de núcleo. Neste trabalho, foi empregado o núcleo de base radial, ou RBF, definido por

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right). \tag{15}$$

Esse núcleo mede a similaridade entre duas observações e permite que a SVM represente fronteiras de decisão não lineares.

Os principais hiperparâmetros do modelo são:

- C: controla o equilíbrio entre uma margem ampla e a penalização dos erros de classificação;
- γ: controla o alcance de influência de cada observação na construção da fronteira de decisão.

Valores elevados de C fazem o modelo penalizar mais intensamente os erros no

treinamento. Valores elevados de γ produzem fronteiras mais localizadas e potencialmente mais complexas.

A SVM pode apresentar elevada capacidade preditiva, mas sua interpretação é menos direta do que a da regressão logística.

4.1.4 Random Forest

A Random Forest é um método de conjunto formado por diversas árvores de decisão. Cada árvore divide os dados sucessivamente de acordo com regras construídas a partir das variáveis preditoras.

As árvores são treinadas em diferentes amostras obtidas por reamostragem com reposição do conjunto de treinamento. Além disso, em cada divisão da árvore, apenas uma parte aleatória das variáveis é considerada.

Esses dois mecanismos aumentam a diversidade entre as árvores e reduzem a dependência do modelo em relação a uma única amostra ou variável. A probabilidade final pode ser calculada pela média das probabilidades estimadas pelas B árvores:

$$\widehat{p}_i = \frac{1}{B} \sum_{b=1}^{B} \widehat{p}_{ib}, \tag{16}$$

em que $\widehat{p}_{ib}$ é a probabilidade estimada pela árvore b para o jogador i.

A Random Forest é capaz de representar relações não lineares e interações entre variáveis sem que essas relações sejam previamente especificadas. Entretanto, o funcionamento conjunto de muitas árvores torna sua interpretação menos direta do que a de um modelo linear.

4.2 Procedimento de validação

A comparação dos modelos foi realizada por validação cruzada aninhada. Esse procedimento utiliza dois ciclos de validação:

- **ciclo interno:** utilizado para selecionar os hiperparâmetros de cada algoritmo;
- **ciclo externo:** utilizado para estimar o desempenho do modelo em observações não utilizadas no ajuste ou na seleção dos hiperparâmetros.

O ciclo externo foi composto por cinco dobras estratificadas. Em cada iteração, quatro dobras foram utilizadas para o desenvolvimento do modelo e uma dobra foi utilizada para sua avaliação.

Dentro dos dados de treinamento de cada ciclo externo, foi realizada uma nova validação cruzada estratificada com quatro dobras para selecionar os hiperparâmetros.

O procedimento pode ser representado de maneira simplificada por

$$\text{ciclo externo}\,[\text{ciclo interno para seleção} \longrightarrow \text{avaliação na dobra externa}]\,. \tag{17}$$

A estratificação manteve aproximadamente a mesma proporção entre jogadores com e sem lesão em todas as dobras.

A validação cruzada aninhada reduz o otimismo que poderia ocorrer caso os mesmos dados fossem utilizados simultaneamente para selecionar e avaliar os modelos.

Após a comparação, o algoritmo selecionado foi ajustado com todas as 640 observações do conjunto de treinamento. A escolha do limiar de classificação também foi realizada exclusivamente nesse conjunto, sem utilizar as observações de teste. Para isso, as probabilidades estimadas foram avaliadas em uma grade de possíveis limiares, e foi selecionado o valor que maximizou a acurácia balanceada no treinamento. Esse critério foi adotado por equilibrar sensibilidade e especificidade, em vez de favorecer apenas uma das classes.

O conjunto de teste, composto por 160 observações, permaneceu completamente isolado durante a seleção do algoritmo, dos hiperparâmetros, do pré-processamento e do limiar, sendo utilizado uma única vez na avaliação final.

Os intervalos de confiança das métricas finais foram estimados por 3.000 reamostragens bootstrap estratificadas.

4.3 Métricas de avaliação

A avaliação considerou métricas relacionadas à discriminação, à classificação e à qualidade das probabilidades estimadas.

Considere os seguintes elementos da matriz de confusão:

- VP: verdadeiros positivos, correspondentes aos jogadores lesionados corretamente classificados;
- VN: verdadeiros negativos, correspondentes aos jogadores não lesionados corretamente classificados;
- FP: falsos positivos, correspondentes aos jogadores sem lesão classificados como lesionados;
- FN: falsos negativos, correspondentes aos jogadores lesionados classificados como não lesionados.

4.3.1 Sensibilidade

A sensibilidade representa a proporção de jogadores lesionados que foram corretamente identificados pelo modelo:

$$\text{Sensibilidade} = \frac{VP}{VP + FN}. \tag{18}$$

Uma sensibilidade elevada indica poucos falsos negativos. Em um contexto de triagem, essa métrica é importante porque um falso negativo corresponde a um jogador lesionado que não foi identificado pelo modelo.

4.3.2 Especificidade

A especificidade representa a proporção de jogadores sem lesão que foram corretamente classificados:

$$\text{Especificidade} = \frac{VN}{VN + FP}. \tag{19}$$

Uma especificidade elevada indica poucos falsos positivos.

4.3.3 Precisão

A precisão, também chamada de valor preditivo positivo, representa a proporção de classificações positivas que correspondem efetivamente a jogadores lesionados:

$$\text{Precisão} = \frac{VP}{VP + FP}. \tag{20}$$

Essa métrica responde à seguinte questão: entre os jogadores classificados como positivos pelo modelo, quantos realmente pertencem à classe positiva?

4.3.4 Acurácia balanceada

A acurácia balanceada corresponde à média entre sensibilidade e especificidade:

$$\text{Acurácia balanceada} = \frac{\text{Sensibilidade} + \text{Especificidade}}{2}. \tag{21}$$

Diferentemente da acurácia convencional, essa métrica atribui a mesma importância às duas classes. Por isso, é especialmente útil quando a distribuição das classes é desigual.

4.3.5 Escore F1

O escore F1 corresponde à média harmônica entre precisão e sensibilidade:

$$F_1 = 2\frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}}. \tag{22}$$

O F1 assume valores entre zero e um e será elevado somente quando precisão e sensibilidade forem simultaneamente elevadas.

4.3.6 Área sob a curva ROC

A curva ROC relaciona a sensibilidade com a taxa de falsos positivos para todos os possíveis limiares de classificação.

A área sob essa curva, denominada ROC-AUC, mede a capacidade do modelo de atribuir probabilidades maiores aos jogadores lesionados do que aos não lesionados.

Se um jogador de cada classe for selecionado aleatoriamente, a ROC-AUC pode ser interpretada como a probabilidade de o modelo atribuir maior risco ao jogador lesionado.

Uma ROC-AUC igual a 0,5 indica desempenho semelhante ao acaso, enquanto valores próximos de 1 indicam elevada capacidade de discriminação.

Como a ROC-AUC considera todos os possíveis limiares, ela não depende da escolha de um único ponto de corte.

4.3.7 Precisão média

A precisão média, ou *average precision*, resume o comportamento da curva precisão-revocação. A revocação corresponde à própria sensibilidade.

De forma simplificada, a métrica pode ser representada por

$$\mathrm{AP} = \sum_k \left(R_k - R_{k-1}\right) P_k, \tag{23}$$

em que P_k representa a precisão e R_k representa a revocação em determinado limiar.

A precisão média atribui maior valor aos modelos que mantêm elevada precisão enquanto identificam uma proporção crescente dos casos positivos. Essa métrica é particularmente informativa em problemas com classes desbalanceadas.

4.3.8 Brier score

O Brier score avalia a diferença quadrática entre as probabilidades estimadas e os desfechos observados:

$$\mathrm{Brier} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{p}_i\right)^2. \tag{24}$$

Diferentemente das métricas baseadas apenas na classificação final, o Brier score considera diretamente as probabilidades produzidas pelo modelo.

Seu valor mínimo é zero, correspondente a previsões probabilísticas perfeitas. Portanto, valores menores indicam melhor desempenho.

4.3.9 Perda logarítmica

A perda logarítmica também avalia as probabilidades estimadas e é definida por

$$\text{LogLoss} = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\widehat{p}_i) + (1-y_i)\log(1-\widehat{p}_i)\right]. \tag{25}$$

Essa métrica penaliza de maneira intensa previsões incorretas realizadas com elevada confiança. Por exemplo, atribuir uma probabilidade próxima de zero a um jogador que pertence à classe positiva produz uma penalização elevada.

Assim como no Brier score, valores menores indicam melhor qualidade das probabilidades.

4.3.10 Critério principal de comparação

A ROC-AUC foi adotada como métrica principal de comparação por avaliar a capacidade discriminativa dos modelos independentemente de um único limiar.

As demais métricas foram utilizadas de maneira complementar. A sensibilidade, a especificidade, a acurácia balanceada e o escore F1 avaliaram o comportamento classificatório, enquanto o Brier score e a perda logarítmica avaliaram a qualidade das probabilidades estimadas.

Essa avaliação conjunta foi necessária porque um modelo pode apresentar boa capacidade de discriminação e, ao mesmo tempo, produzir probabilidades inadequadamente calibradas.

4.4 Avaliação e comparação dos modelos

A Tabela 3 apresenta os resultados médios da validação cruzada aninhada.

Tabela 3: Desempenho médio dos modelos na validação cruzada aninhada.

Modelo	**ROC-AUC**	**DP**	**AP**	**Acurácia bal.**	**Brier**
Regressão logística	0,9892	0,0036	0,9895	0,9359	0,0533
SVM com núcleo RBF	0,9887	0,0046	0,9889	0,9422	0,0422
Random Forest	0,9821	0,0058	0,9815	0,9344	0,0721
Linha de base	0,5000	0,0000	0,5000	0,5000	0,2500

Todos os modelos treinados apresentaram desempenho superior à linha de base. A regressão logística obteve o maior ROC-AUC médio, seguida pela SVM e pela Random Forest.

Considerando a ROC-AUC como critério principal, a regressão logística apresentou o maior desempenho médio na validação cruzada aninhada. A SVM obteve valores ligeiramente superiores em algumas métricas complementares, como acurácia balanceada e Brier score, mas a diferença de discriminação entre os modelos foi pequena. Assim, a regressão logística foi selecionada por combinar elevado desempenho preditivo com maior transparência interpretativa.

A especificação final utilizou regularização L2 com parâmetro

$$C = 0{,}01. \tag{26}$$

Além do hiperparâmetro de regularização, foi escolhido um limiar de decisão para transformar as probabilidades estimadas em classes. O valor selecionado foi

$$\tau = 0{,}45. \tag{27}$$

Assim, o limiar de 0,45 não corresponde ao ponto de corte padrão de 0,5, nem foi adotado de forma arbitrária. Ele representa o limiar ótimo obtido no conjunto de treinamento segundo o critério de maior acurácia balanceada. Essa escolha é adequada para avaliar o modelo no contexto deste estudo, mas não deve ser interpretada como um valor universal. Em uma aplicação externa, o limiar deveria ser reavaliado de acordo com a prevalência observada, a calibração das probabilidades e o custo relativo de falsos positivos e falsos negativos.

A análise de sensibilidade mostrou ROC-AUC médio de 0,9892 para a especificação com altura e massa corporal e de 0,9889 para a especificação alternativa com IMC. Assim, a representação antropométrica exerceu impacto mínimo sobre a discriminação.

4.5 Resultados no conjunto de teste

Após a seleção do algoritmo, dos hiperparâmetros e do limiar, o modelo foi ajustado com as 640 observações de treinamento e aplicado uma única vez às 160 observações do teste.

A Tabela 4 apresenta os resultados finais.

Tabela 4: Desempenho da regressão logística no conjunto de teste.

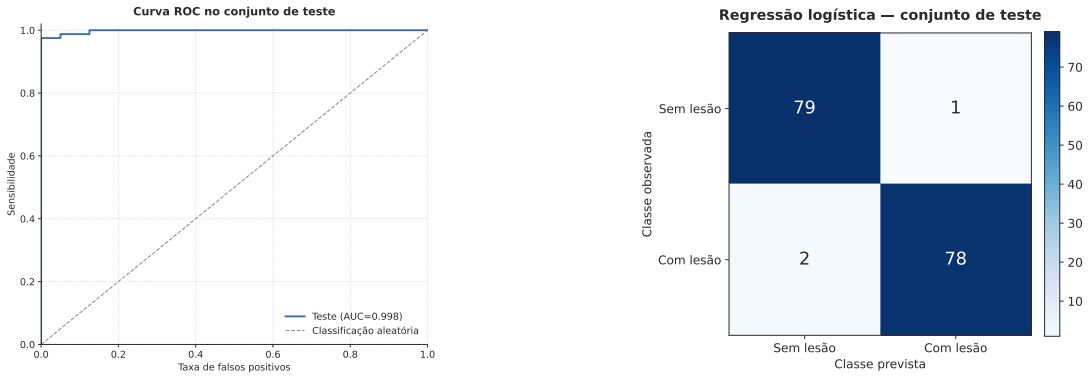
Métrica	**Estimativa**	**IC 95% inferior**	**IC 95% superior**
ROC-AUC	0,9978	0,9934	1,0000
Precisão média	0,9980	0,9941	1,0000
Acurácia balanceada	0,9813	0,9625	1,0000
Sensibilidade	0,9750	0,9375	1,0000
Especificidade	0,9875	0,9625	1,0000
Escore F1	0,9811	0,9610	1,0000
Brier score	0,0570	0,0472	0,0674

Os intervalos de confiança foram estimados por 3.000 reamostragens bootstrap estratificadas.

A matriz de confusão apresentou:

- 79 verdadeiros negativos;
- 1 falso positivo;
- 2 falsos negativos;
- 78 verdadeiros positivos.

Portanto, o modelo classificou corretamente 157 das 160 observações.



(a) Curva ROC. **(b)** Matriz de confusão.

Figura 2: Avaliação da regressão logística no conjunto de teste.

A discriminação observada no teste foi muito elevada. Entretanto, o Brier score passou de 0,0533 durante o desenvolvimento do modelo para 0,0570 no teste. Isso mostra que elevada capacidade de ordenação e classificação não implica probabilidades perfeitamente calibradas.

4.6 Interpretação do modelo

Como as variáveis numéricas foram padronizadas, seus coeficientes representam a mudança no logaritmo das chances associada ao aumento de um desvio padrão.

A Figura 3 apresenta os coeficientes do modelo final.

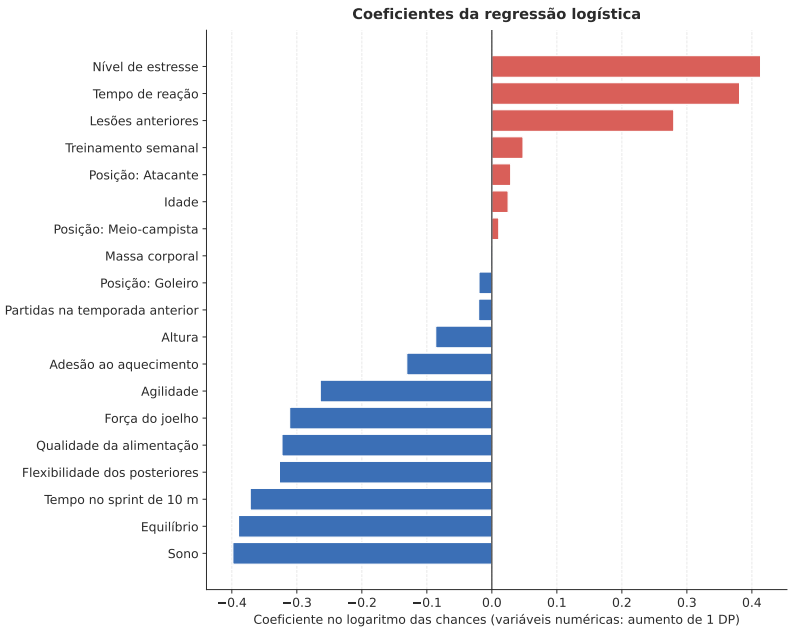


Figura 3: Coeficientes padronizados da regressão logística.

As dez características de maior influência são apresentadas na Tabela 5.

Tabela 5: Principais coeficientes da regressão logística.

Característica	β	**OR**	**Estabilidade**
Nível de estresse	0,413	1,512	100,0%
Sono	-0,398	0,671	100,0%
Equilíbrio	-0,390	0,677	100,0%
Tempo de reação	0,381	1,464	100,0%
Velocidade no sprint de 10 m	-0,372	0,690	100,0%
Flexibilidade	-0,327	0,721	100,0%
Qualidade da alimentação	-0,323	0,724	100,0%
Força do joelho	-0,311	0,733	100,0%
Lesões anteriores	0,280	1,323	100,0%
Agilidade	-0,264	0,768	100,0%

Estresse, tempo de reação e histórico de lesões apresentaram associações positivas com as probabilidades estimadas. Sono, equilíbrio, desempenho no sprint de 10 metros, flexibilidade, qualidade da alimentação, força do joelho e agilidade apresentaram

associações negativas.

A variável `Sprint_Speed_10m_s` foi interpretada como um indicador de desempenho ou velocidade no sprint de 10 metros. Portanto, o coeficiente negativo observado sugere que maiores valores desse indicador estiveram associados a menores probabilidades estimadas de lesão, mantendo constantes as demais variáveis. Esse resultado não deve ser interpretado como evidência causal de que maior velocidade previne lesões. O desempenho no sprint pode funcionar como marcador de aptidão neuromuscular, preparação física ou capacidade de tolerar ações de alta intensidade. Por outro lado, em contextos reais, a exposição frequente a corridas de alta velocidade também pode aumentar o risco quando não acompanhada de adaptação adequada. Assim, a associação observada deve ser entendida como específica da base analisada.

A estabilidade das interpretações foi avaliada por 1.000 reamostragens bootstrap. Estresse, sono, equilíbrio, tempo de reação e desempenho no sprint mantiveram o mesmo sinal em todas as reamostragens e apareceram com frequência entre os fatores de maior magnitude.

A adesão ao aquecimento apresentou razão de chances ajustada igual a

$$\mathrm{OR} = 0{,}877. \tag{28}$$

Após o controle pelas demais variáveis, a adesão esteve associada a chances estimadas aproximadamente 12,3% menores. A diferença em relação à razão de chances não ajustada demonstra que parte da associação foi compartilhada com outros preditores.

4.7 Discussão e limitações

A regressão logística apresentou desempenho comparável aos modelos não lineares e ofereceu maior transparência. Dessa forma, a seleção do modelo final não foi baseada apenas na maior métrica possível, mas no equilíbrio entre desempenho e interpretabilidade.

As variáveis de maior influência pertencem a diferentes dimensões do risco de lesão. Estresse e sono representam aspectos comportamentais e de recuperação, enquanto equilíbrio, força, flexibilidade, agilidade, reação e sprint representam capacidades

funcionais. Esse resultado reforça a necessidade de uma abordagem multivariada.

Apesar disso, o desempenho excepcionalmente elevado deve ser interpretado com cautela. A base possui classes perfeitamente equilibradas, não apresenta valores ausentes e contém forte separação entre vários preditores e o desfecho. A proporção exata de 400 jogadores com lesão e 400 sem lesão foge do que normalmente se espera em uma amostra observacional não controlada e sugere que o conjunto pode ter sido curado, balanceado ou construído com finalidade didática/preditiva. Esse cenário é mais simples do que aquele normalmente encontrado em dados reais de clubes, nos quais a incidência de lesão dependeria da exposição e da definição clínica adotada.

Além disso, o limiar final de classificação foi escolhido de forma otimizada no conjunto de treinamento. Essa decisão melhora a adequação entre sensibilidade e especificidade na base analisada, mas também reforça que o ponto de corte obtido é específico deste conjunto de dados. Em outra população, a distribuição do desfecho, a calibração do modelo e os custos práticos de cada tipo de erro poderiam levar a um limiar diferente.

A avaliação realizada representa validação interna. Embora o conjunto de teste não tenha sido utilizado durante o treinamento, todas as observações possuem a mesma origem. Portanto, os resultados não demonstram generalização para outras equipes, categorias, países ou temporadas.

Também não estão integralmente documentados os instrumentos utilizados para construir todos os escores, o processo de seleção dos jogadores e os critérios clínicos detalhados empregados na definição do desfecho. A base também não informa tipo, localização, gravidade, tempo de afastamento ou recorrência das lesões.

Outra limitação é o caráter predominantemente estático das variáveis. Modelos dinâmicos de etiologia ressaltam que o risco de lesão se modifica ao longo do tempo conforme exposição, recuperação, estado clínico e eventos esportivos [9]. Em uma aplicação real, seria desejável utilizar medidas repetidas de carga interna e externa, dor, fadiga, sono, estresse, minutos de exposição e estado clínico, integrando diferentes indicadores de resposta ao treinamento [1, 10].

O modelo não deve ser utilizado automaticamente para afastar atletas ou restringir sua participação. Sua possível função seria apoiar a triagem e indicar jogadores

que poderiam receber avaliação adicional. Qualquer uso dependeria de validação externa e integração com o julgamento dos profissionais responsáveis.

5 Conclusão

Este trabalho desenvolveu uma abordagem interpretável de aprendizado de máquina para estimar a ocorrência de lesão na temporada seguinte em jogadores universitários de futebol. A proposta consistiu em comparar diferentes modelos de classificação, avaliar seu desempenho preditivo e interpretar as variáveis mais associadas às estimativas produzidas.

A regressão logística regularizada foi selecionada por apresentar desempenho comparável ao dos modelos não lineares e, ao mesmo tempo, maior transparência na interpretação dos resultados. No conjunto de teste, o modelo obteve ROC-AUC de 0,9978, sensibilidade de 0,9750, especificidade de 0,9875 e acurácia balanceada de 0,9813.

As características de maior influência e estabilidade foram estresse, sono, equilíbrio, tempo de reação e velocidade no sprint de 10 metros. Histórico de lesões, flexibilidade, força do joelho, agilidade, qualidade da alimentação e adesão ao aquecimento também contribuíram de maneira consistente para as estimativas do modelo.

Entretanto, os resultados devem ser interpretados com cautela. As associações identificadas não demonstram relações causais, nem permitem afirmar que as variáveis mais influentes sejam, isoladamente, causas ou formas diretas de prevenção das lesões. Além disso, o desempenho elevado foi obtido em uma única base pública, limpa, sem valores ausentes, perfeitamente equilibrada entre as classes e sem validação externa. A proporção de 50% de jogadores lesionados deve ser entendida como uma característica da base analisada, e não como uma estimativa da incidência real de lesões no futebol.

O limiar de decisão também deve ser interpretado dentro desse contexto. O valor $\tau = 0{,}45$ foi selecionado como limiar ótimo no conjunto de treinamento, com o objetivo de maximizar a acurácia balanceada, e não representa necessariamente o melhor ponto de corte para outras populações. Em aplicações reais, o limiar precisaria ser recalibrado considerando a prevalência observada, a qualidade das probabilidades estimadas e as consequências práticas de falsos positivos e falsos negativos.

Por essas razões, o estudo deve ser entendido como uma prova de conceito metodológica, e não como uma estimativa epidemiológica direta do risco de lesão ou como uma ferramenta pronta para uso clínico.

A ausência de registros longitudinais também limitou a incorporação de medidas de *training load*, como carga semanal acumulada, carga aguda, carga crônica ou razão carga aguda:crônica. A inclusão desse tipo de informação permitiria comparar as probabilidades estimadas com indicadores de exposição recente, variações abruptas de carga e padrões individuais de adaptação ao treinamento.

Como continuidade, recomenda-se realizar coleta prospectiva em múltiplas temporadas, utilizar definições padronizadas de lesão, incorporar dados longitudinais de exposição e carga de treinamento e validar o modelo em jogadores provenientes de outras populações.

Conclui-se que modelos multivariados e interpretáveis podem organizar informações heterogêneas e identificar padrões associados à ocorrência de lesões. Seu valor prático, contudo, dependerá da qualidade dos dados, da validação em novos contextos e da utilização responsável em conjunto com a avaliação de profissionais do esporte e da saúde.

Configurações da modelagem

Tabela 6: Principais configurações utilizadas na modelagem.

Configuração	**Valor utilizado**
Proporção destinada ao conjunto de teste	20%
Semente aleatória	42
Método de divisão dos dados	Divisão estratificada
Validação cruzada externa	5 dobras
Validação cruzada interna	4 dobras
Modelo final	Regressão logística
Tipo de regularização	L2
Parâmetro de regularização	$C = 0{,}01$
Limiar de classificação	$\tau = 0{,}45$
Critério de escolha do limiar	Maior acurácia balanceada no conjunto de treinamento
Bootstrap dos coeficientes	1.000 reamostragens
Bootstrap das métricas finais	3.000 reamostragens

Referências

[1] Pitre C. Bourdon, Marco Cardinale, Andrew Murray, Paul Gastin, Michael Kellmann, Matthew C. Varley, Tim J. Gabbett, Aaron J. Coutts, Darren J. Burgess, Warren Gregson, and N. Timothy Cable. Monitoring athlete training loads: Consensus statement. *International Journal of Sports Physiology and Performance*, 12 (Suppl 2):S2–161–S2–170, 2017. doi: 10.1123/IJSPP.2017-0208.

[2] Filipe Manuel Clemente, José Afonso, Júlio Costa, Rafael Oliveira, José Pino-Ortega, and Markel Rico-González. Relationships between sleep, athletic and match performance, training load, and injuries: A systematic review of soccer players. *Healthcare*, 9(7):808, 2021. doi: 10.3390/healthcare9070808.

[3] Jan Ekstrand, Martin Hägglund, and Markus Waldén. Injury incidence and injury patterns in professional football: the uefa injury study. *British Journal of Sports Medicine*, 45(7):553–558, 2011. doi: 10.1136/bjsm.2009.060582.

[4] Carolyn A. Emery and Willem H. Meeuwisse. The effectiveness of a neuromuscular prevention strategy to reduce injuries in youth soccer: a cluster-randomised controlled trial. *British Journal of Sports Medicine*, 44(8):555–562, 2010. doi: 10.1136/bjsm.2010.074377.

[5] Martin Hägglund, Markus Waldén, and Jan Ekstrand. Previous injury as a risk factor for injury in elite football: a prospective study over two consecutive seasons. *British Journal of Sports Medicine*, 40(9):767–772, 2006. doi: 10.1136/bjsm.2006.026609.

[6] Andreas Ivarsson, Urban Johnson, Mark B. Andersen, Ulrika Tranaeus, Andreas Stenling, and Magnus Lindwall. Psychosocial factors and sport injuries: Meta-analyses for prediction and prevention. *Sports Medicine*, 47:353–365, 2017. doi: 10.1007/s40279-016-0578-x.

[7] Jiacheng Ma, Shengrui Liu, and Yuting Pei. Shap-based interpretable machine learning for injury risk prediction in university football players: a multi-dimensional data analysis approach. *Scientific Reports*, 15:40252, 2025. doi: 10.1038/s41598-025-24144-y.

[8] Aritra Majumdar, Rashid Bakirov, Dan Hodges, Suzanne Scott, and Tim Rees. Machine learning for understanding and predicting injuries in football. *Sports Medicine - Open*, 8(1):73, 2022. doi: 10.1186/s40798-022-00465-4.

[9] Willem H. Meeuwisse, H. Tyreman, Brent Hagel, and Carolyn Emery. A dynamic model of etiology in sport injury: the recursive nature of risk and causation. *Clinical Journal of Sport Medicine*, 17(3):215–219, 2007. doi: 10.1097/JSM.0b013e3180592a48.

[10] Torbjørn Soligard, Martin Schwellnus, Juan-Manuel Alonso, Roald Bahr, Benjamin Clarsen, H. Paul Dijkstra, Tim Gabbett, Michael Gleeson, Martin Hägglund, Mark R. Hutchinson, Christa Janse van Rensburg, Karim M. Khan, Romain Meeusen, John W. Orchard, Babette M. Pluim, Martin Raftery, Richard Budgett, and Lars Engebretsen. How much is too much? (part 1) international olympic committee consensus statement on load in sport and risk of injury. *British Journal of Sports Medicine*, 50(17):1030–1041, 2016. doi: 10.1136/bjsports-2016-096581.

[11] Kjell Svensson, Marie Alricsson, Marcus Olausson, and Suzanne Werner. Physical performance tests – a relationship of risk factors for muscle injuries in elite level male football players. *Journal of Exercise Rehabilitation*, 14(2):282–288, 2018. doi: 10.12965/jer.1836028.014.

[12] Kristian Thorborg, Kasper Kühn Krommes, Ernest Esteve, Mikkel Bek Clausen, Else Marie Bartels, and Michael Skovdal Rathleff. Effect of specific exercise-based football injury prevention programmes on the overall injury rate in football: a systematic review and meta-analysis of the fifa 11 and 11+ programmes. *British Journal of Sports Medicine*, 51(7):562–571, 2017. doi: 10.1136/bjsports-2016-097066.

[13] Yuan Chunhong. University Football Injury Prediction Dataset. Kaggle, 2025. URL `https://www.kaggle.com/datasets/yuanchunhong/university-football-injury-predict` Acesso em: 06 jun. 2026.