

PROPOSTA DE COLABORAÇÃO — TESE DE DOUTORAMENTO

# Predição de Substituições no Futebol com Machine Learning

---

Plataforma operacional com modelos ML validados  
Pronta para validação científica com dados reais de jogo

Plataforma Operacional

XGBoost + SHAP

Pronto para Validação

Rui Pedro Ribeiro dos Santos  
Universidade de Trás-os-Montes e Alto Douro

# Objetivo da Tese de Doutoramento

Desenvolver e validar um modelo preditivo baseado em Machine Learning capaz de recomendar o **momento ótimo de substituição** ao longo de um jogo de futebol, integrando múltiplas fontes de dados.

## Pergunta de Investigação

*"É possível prever, com base em dados de GPS, carga de treino, bem-estar e análise de vídeo, qual o momento ótimo para efetuar substituições durante um jogo, maximizando a performance coletiva?"*

## Hipótese Central

A integração de dados multi-fonte com modelos de ML explicáveis (XGBoost + SHAP) permite identificar padrões de queda de performance em tempo real, antecipando a necessidade de substituição antes que o rendimento se deteriore significativamente.

### ✓ Fase 1 — Concluída

Plataforma web completa desenvolvida (FastAPI + React) com integração de dados GPS, PSE, Wellness e Vídeo

### ✓ Fase 2 — Concluída

Modelo ML de predição pré-jogo implementado e operacional (XGBoost + SHAP) com 59 features e explicabilidade completa

### ► Fase 3 — Validação Científica

Recolha de dados de uma época completa para validar modelo preditivo — **necessita parceria com clube**

### ► Fase 4 — Publicação

Análise estatística, validação dos resultados e publicação científica em revistas de alto impacto

# As substituições são ainda decisões subjetivas

---

Apesar da enorme quantidade de dados disponíveis no futebol moderno, a decisão de quando substituir um jogador continua a ser maioritariamente baseada na intuição do treinador.

## Dados em silos separados

GPS num software, wellness numa folha, vídeo noutra — sem integração

## Sem modelos preditivos para substituições

A literatura foca-se em prevenção de lesões, não no timing de substituição

## Substituições tardias custam pontos

Estudos mostram que substituições no momento errado afetam o resultado final



## O que falta na investigação

1. Um modelo que integre dados pré-jogo (carga semanal, wellness, risco) com dados em tempo real (GPS ao vivo) para prever queda de performance durante o jogo
2. Explicabilidade — o treinador precisa de saber *porquê*, não apenas *quando*
3. Validação com dados reais de jogos profissionais ao longo de uma época

## TRABALHO REALIZADO

# Plataforma 100% Operacional

Sistema completo desenvolvido e testado. Plataforma web profissional com modelos ML validados em dados simulados, pronta para deployment e validação científica com dados reais.



### Ingestão Multi-fonte

GPS (Catapult CSV), questionários PSE/Bem-estar, vídeo de treino e jogo



### Métricas Avançadas

ACWR, Monotonia, Tensão, Z-scores, Risk Assessment multi-fatorial



### ML + Explicabilidade

XGBoost com 26 features e SHAP para predição de queda de performance



### Computer Vision

YOLOv8 para deteção de jogadores e bola em vídeo de jogo



### Avaliação de Risco

Risco de lesão, performance e substituição com recomendações automáticas



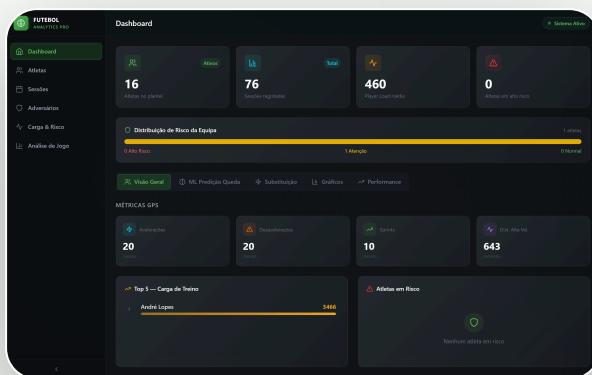
### Dashboard Interativo

Interface profissional para staff técnico com visualizações em tempo real

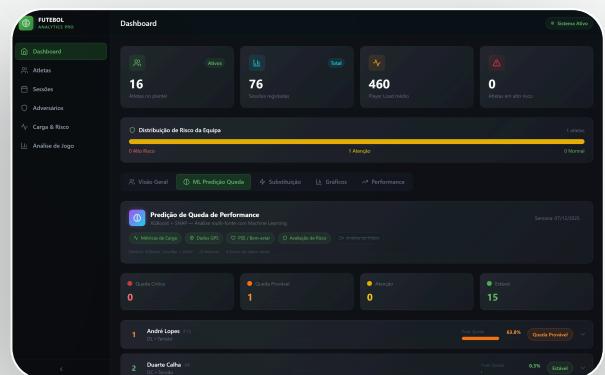
## PLATAFORMA DESENVOLVIDA

# Dashboard em funcionamento

## VISÃO GERAL — Métricas GPS, Risco, Carga



## ML — Previsão de Queda com SHAP



**26**

Features no  
Modelo ML

**6**

Categorias de  
Dados

**5**

Fontes  
Integradas

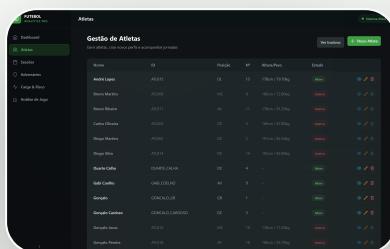
**100%**

Explicabilidade SHAP

## MÓDULOS DA PLATAFORMA

# Recolha e análise integrada

### GESTÃO DE ATLETAS



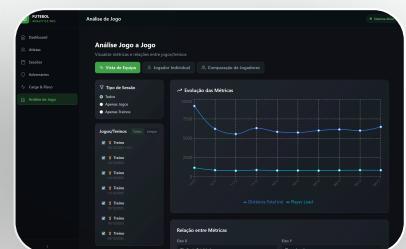
Perfis completos, dados físicos, estado e histórico de sessões por jogador.

### CARGA & RISCO



Monotonia, tensão e ACWR por posição. Classificação automática de risco.

### ANÁLISE DE JOGO



Evolução de métricas ao longo dos jogos/treinos com comparação entre jogadores.

# Machine Learning para Predição de Substituições

## Dados de Entrada (por jogador, por jogo)

FONTE	DADOS	TIMING
GPS	Distância, sprints, acelerações, player load	Tempo real
Wellness	Fadiga, sono, dor muscular, wellness score	Pré-jogo
Carga	ACWR, monotonia, tensão, carga semanal	Pré-jogo
Risco	Risco lesão, fadiga acumulada, wellness trend	Pré-jogo
Vídeo	Posicionamento, movimentação, padrões táticos	Tempo real

Variável alvo: Momento ótimo de substituição =  $f(\text{estado\_pré\_jogo}, \text{performance\_em\_jogo}, \text{contexto\_tático})$   
 Modelo: XGBoost + SHAP TreeExplainer  
 Validação: K-fold CV + validação temporal

## Pipeline de Predição

### 1. Estado Pré-Jogo

Carga acumulada, wellness, risco de lesão, histórico recente

### 2. Monitorização Em Jogo

GPS em tempo real: distância, sprints, intensidade por período

### 3. Detecção de Queda

ML identifica padrões de declínio comparando com baseline do jogador

### 4. Recomendação Explicável

SHAP mostra ao treinador *porque* substituir e *quando*

# Para validar o modelo, preciso de dados reais

---

A plataforma está construída e funcional. O modelo de predição de queda de performance já funciona com dados de treino. Para estender à **predição de substituições em jogo**, preciso de dados reais de competição.

## Dados necessários do clube:



**Exports GPS de jogos e treinos**  
Catapult/STATSports CSV — já suportado pela plataforma



**Questionários de bem-estar / PSE**  
Fadiga, sono, dor muscular — diários ou pré-sessão



**Registros de substituições em jogo**  
Minuto, jogador substituído, jogador que entrou, contexto



**Vídeo de jogos (opcional)**  
Para análise de computer vision — posicionamento e padrões táticos

**Duração estimada da recolha**

**1 Época**

Período ideal de recolha de dados

**Mínimo viável:** ~20 jogos oficiais com dados GPS completos e registos de substituições

**Ideal:** Época completa (~40 jogos) + treinos semanais + wellness diário

**Impacto zero no dia-a-dia do clube**

A plataforma consome dados que o clube já recolhe (GPS, wellness). Não requer mudanças nos processos existentes — apenas acesso aos exports.

# O que o clube **ganha** com esta colaboração



## Acesso gratuito à plataforma

Dashboard completo de monitorização de performance, carga e risco — sem custos para o clube durante toda a colaboração.



## Predições ML personalizadas

Modelo treinado especificamente com os dados do clube — predições de queda de performance e recomendações de substituição.



## Relatórios de investigação

Acesso aos resultados da investigação: padrões de performance, fatores de risco e insights específicos do plantel.



## Confidencialidade total

Dados anonimizados na publicação. Acordo de confidencialidade formal. O clube controla o que é partilhado.



## Inovação científica

Associação do clube a investigação de ponta em ciências do desporto e inteligência artificial aplicada ao futebol.

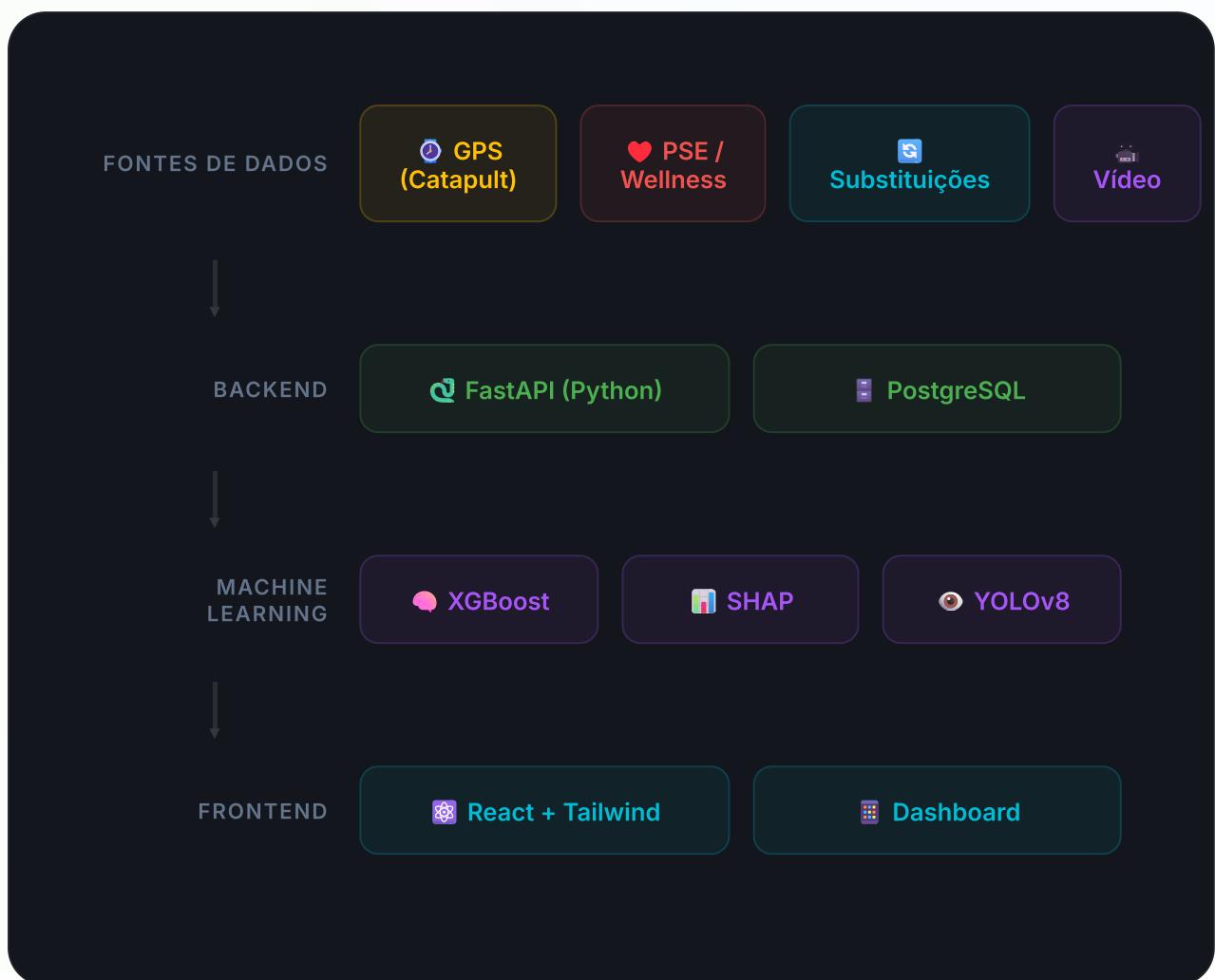


## Suporte técnico contínuo

Manutenção e evolução da plataforma durante toda a colaboração, adaptada às necessidades do staff técnico.

ARQUITETURA TÉCNICA

# Stack Tecnológico



# Baseado em investigação

## ACWR — Acute:Chronic Workload Ratio

Gabbett, T.J. (2016). The training-injury prevention paradox. *British Journal of Sports Medicine*.

## Monitorização de Carga de Treino

Bourdon, P.C. et al. (2017). Monitoring athlete training loads. *Int. J. Sports Physiology and Performance*.

## Substituições e Performance

Bradley, P.S. et al. (2014). The effect of playing formation on high-intensity running and technical profiles. *J. Sports Sciences*.

## SHAP — Explicabilidade em ML

Lundberg, S.M. & Lee, S.I. (2017). A unified approach to interpreting model predictions. *NeurIPS*.

## XGBoost em Dados Desportivos

Bunker, R.P. & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*.

## Metodologia de Investigação

- ✓ Revisão sistemática da literatura
- ✓ Desenvolvimento da plataforma de recolha
- ✓ Feature engineering baseado em evidência
- ✓ Validação cruzada temporal (K-fold CV)
- ✓ Explicabilidade com SHAP values
- ✓ Validação ecológica com dados reais de jogo

## Publicações Previstas

- ✓ Artigo 1: Plataforma de integração multi-fonte
- ✓ Artigo 2: Modelo de predição de substituições
- ✓ Artigo 3: Validação com dados de época completa
- ✓ Tese de doutoramento

# Proteção de dados **garantida**



## Acordo de Confidencialidade

Assinatura de NDA formal antes do início da colaboração. O clube aprova toda a informação antes de qualquer publicação.



## Anonimização

Todos os dados publicados são anonimizados. Nomes de jogadores, staff e clube são codificados nas publicações.



## Comissão de Ética

Estudo aprovado pela comissão de ética da universidade. Consentimento informado de todos os participantes.

## O clube mantém controlo total sobre:

- ✓ Quais dados são partilhados com o investigador
- ✓ Aprovação prévia de qualquer publicação
- ✓ Direito de retirada a qualquer momento

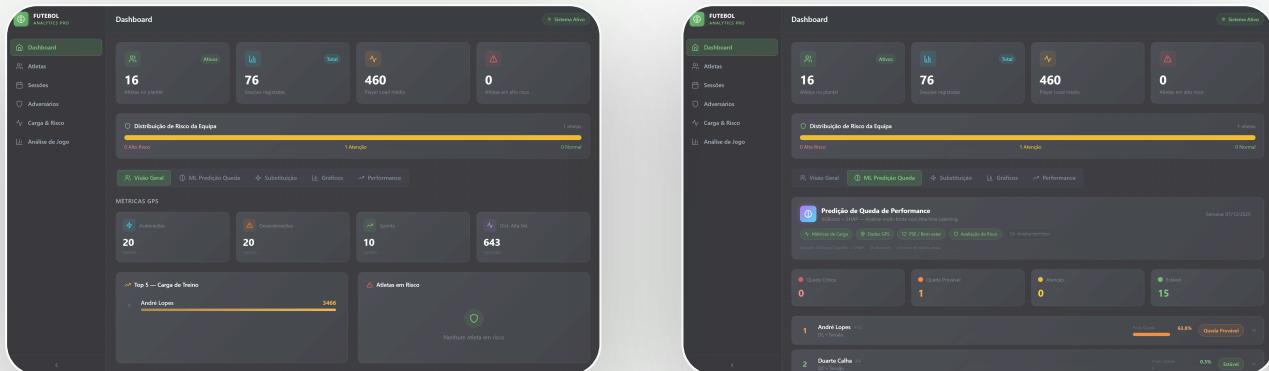
## O investigador compromete-se a:

- ✓ Não partilhar dados com terceiros
- ✓ Eliminar dados brutos após conclusão do estudo
- ✓ Disponibilizar todos os resultados ao clube

## DEMONSTRAÇÃO

# Demonstração da Plataforma

Plataforma funcional com dados de teste — pronta para receber dados reais do clube.



## PROPOSTA DE COLABORAÇÃO

# Vamos construir isto juntos?

A plataforma está pronta. O modelo está implementado.  
Para dar o próximo passo — prever substituições em jogo —  
**preciso da vossa colaboração.**



### Dados GPS

Exports de jogos  
e treinos



### Wellness

Questionários  
diários



### Substituições

Registos de jogo



### Vídeo

Opcional

Custo zero para o clube

Confidencialidade total

Plataforma gratuita

Resultados partilhados

Rui Pedro Ribeiro dos Santos

Universidade de Trás-os-Montes e Alto Douro

✉ al64943@alunos.utad.pt · rui.coach.performance@gmail.com

☎ +351 932 177 858

Obrigado pela atenção.