UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Previsão e Análise de Lesões em Corredores: Um Estudo Preditivo Baseado em Dados de Treinamento

Juliano Rodrigues Dourado

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Juliano Rodrigues Dourado

Previsão e Análise de Lesões em Corredores: Um Estudo Preditivo Baseado em Dados de Treinamento

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Marcelo Manzato

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m	Dourado, Juliano Rodrigues Previsão e Análise de Lesões em Corredores: Um Estudo Preditivo Baseado em Dados de Treinamento / Juliano Rodrigues Dourado ; orientador Marcelo Manzato. – São Carlos, 2025. 46 p. : il. (algumas color.) ; 30 cm. Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025. 1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Manzato, Marcelo Garcia, orient. II. Título.

Juliano Rodrigues Dourado

Injury Prediction and Analysis in Runners: A Predictive Study Based on Training Data

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Marcelo Manzato

Original version

São Carlos

2025

Dedico este trabalho à minha família — meus pais, irmãos, esposa e sogros — por todo o apoio, incentivo e carinho ao longo desta jornada. A cada um de vocês, minha gratidão por acreditarem em mim

AGRADECIMENTOS

Agradeço, em primeiro lugar à minha esposa Janaína, por todo o apoio, paciência e encorajamento ao longo desta jornada. Sua parceria foi essencial para que eu pudesse conciliar os desafios pessoais, profissionais e acadêmicos.

Agradeço também ao meu orientador, Prof. Dr. Marcelo Manzato, pela orientação atenciosa, pelos conselhos técnicos e pela disponibilidade em todas as etapas deste trabalho. Sua contribuição foi decisiva para a qualidade e a conclusão deste estudo.

Aos meus colegas da Turma IV do MBA em Inteligência Artificial e Big Data, deixo meu reconhecimento pelas trocas de conhecimento, pelas conversas enriquecedoras e pela convivência ao longo do curso.

Por fim, agradeço às ferramentas de Inteligência Artificial que foram grandes aliadas na revisão do texto deste trabalho.

"Aprender é a única coisa de que a mente nunca se cansa, nunca tem medo e nunca se arrependе."

Leonardo da Vinci

RESUMO

Dourado, J. **Previsão e Análise de Lesões em Corredores: Um Estudo Preditivo Baseado em Dados de Treinamento**. 2025. 46 p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Este estudo apresenta um modelo preditivo baseado em aprendizado de máquina, incluindo técnicas de aprendizado profundo para identificar riscos de lesões em corredores de alto rendimento, utilizando dados temporais de treinamento. Foram analisados dois conjuntos públicos (diários e semanais) com informações de 74 atletas holandeses (2012–2019), incluindo quilometragem, intensidade, esforço percebido e recuperação. A metodologia envolveu o pré-processamento dos dados, a criação de features temporais (como médias móveis e variações) e o desenvolvimento de modelos individuais (Regressão Logística, Random Forest, XGBoost, LSTM), além de um modelo Ensemble que combina os melhores resultados. Para lidar com o desbalanceamento das classes (1,34% de lesões), foi aplicada validação cruzada estratificada, avaliando métricas como Recall, F2-Score, Balanced Accuracy e AUC-ROC. Os resultados indicam que o modelo Ensemble superou os individuais, alcançando F2-Score de 0,1773 e AUC-ROC de 0,6926, com maior precisão e redução de falsos positivos, mesmo com recall moderado. Métricas relacionadas ao volume, intensidade e recuperação foram identificadas como preditores-chave, confirmando a eficácia da engenharia temporal. Na prática, o modelo sugere que monitorar tendências de treino, manter a consistência e controlar a intensidade são estratégias eficazes para prevenir lesões.

Palavras-chave: Atletas de Corrida. Prevenção de Lesões. Carga de Treinamento. Modelos Ensemble. Dados Temporais.

ABSTRACT

Dourado, J. **Injury Prediction and Analysis in Runners: A Predictive Study Based on Training Data**. 2025. 46 p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

This study presents a machine learning-based predictive model to identify injury risks in high-performance runners using temporal training data. Two public datasets (daily and weekly) from 74 Dutch athletes (2012–2019) were analyzed, including metrics such as mileage, intensity, perceived effort, and recovery. The methodology involved data preprocessing, temporal feature engineering (e.g., moving averages and variations), and the development of individual models (Logistic Regression, Random Forest, XGBoost, LSTM), as well as an Ensemble model combining their outputs. To address class imbalance (1,34% injury cases), stratified cross-validation was applied, evaluating metrics such as Recall, F2-Score, Balanced Accuracy, and AUC-ROC. Results indicate that the Ensemble model outperformed individual models, achieving an F2-Score of 0.1773 and an AUC-ROC of 0.6926, with improved precision and reduced false positives, despite moderate recall. Key predictors related to training volume, intensity, and recovery confirmed the effectiveness of temporal feature engineering. Practically, the model suggests that monitoring training trends, maintaining consistency and managing intensity are effective strategies for injury prevention.

Keywords: Running Athletes, Injury Prevention, Training Load, Ensemble Models, Temporal Data.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

F2-Score	*F-beta Score with β=2* - Métrica de avaliação que dá maior peso ao recall do que à precisão
AUC-ROC	*Area Under the Receiver Operating Characteristic Curve*
FCM	Frequência Cardíaca Máxima
VO_2máx	Consumo máximo de oxigênio – indicador da capacidade aeróbica
ACWR	*Acute:Chronic Workload Ratio*
AUC-PR	*Area Under the Precision-Recall Curve*
LSTM	*Long Short-Term Memory*
MI	*Mutual Information*
RF	*Random Forest*
SMOTE	*Synthetic Minority Over-sampling Technique*
BPM	*Batimentos por minuto*
XGBoost	*Extreme Gradient Boosting*
Z3-4	Zona de Intensidade 3-4 (limiar anaeróbico)
Z5-T1-T2	Zona de Intensidade 5, *Threshold*Threshold 1-2 (alta intensidade)

SUMÁRIO

1 INTRODUÇÃO

A corrida de rua tem conquistado cada vez mais adeptos entre os atletas amadores, seja como uma maneira de manter a motivação e a regularidade nos treinos, seja como uma oportunidade de competir e alcançar resultados. No entanto, manter uma rotina consistente de treinos visando melhorar a performance e ao mesmo tempo minimizar o risco de lesões é um desafio comum entre atletas de todos os níveis. O treinamento de corredores, especialmente de médias e longas distâncias, envolve aspectos físicos, psicológicos e nutricionais, que podem impactar significativamente o desempenho e a recuperação, como apontado por (Zinner; Sperlich, 2015). Apesar disso, as lesões continuam sendo um problema recorrente, com estudos indicando que até 50% dos praticantes enfrentam algum tipo de lesão.

Com o avanço das tecnologias de rastreamento de dados, como dispositivos de monitoramento de desempenho e plataformas digitais, há uma grande quantidade de informações que, se bem analisadas, podem fornecer insights sobre os fatores de risco para lesões. No entanto, poucos modelos computacionais exploram essas informações de forma eficaz. A análise de dados como carga acumulada, intensidade e variabilidade dos treinos pode ser uma ferramenta importante para identificar padrões que levam a lesões. Diante disso, este trabalho explora a previsão de lesões em corredores, utilizando exclusivamente dados de treinamento, como intensidade, carga acumulada e outros parâmetros registrados durante as atividades físicas, com o objetivo de desenvolver um sistema para prever lesões e propor estratégias preventivas.

A escolha deste tema justifica-se pela alta prevalência de lesões e pela crescente demanda por métodos para mitigá-las, promovendo um treinamento mais seguro e eficaz. A análise dos dados pode também contribuir para entender os limites físicos e personalizar os treinos, ajudando os atletas a evitar sobrecarga e otimizar o desempenho.

O objetivo deste trabalho é desenvolver um sistema para prever lesões em corredores, utilizando o conjunto de dados *"Injury Prediction in Competitive Runners with Machine Learning"*, (Lovdal; Hartigh; Azzopardi, 2020), composto por um registro detalhado de treinamentos de uma equipe holandesa de corrida de alto nível, com 74 atletas, coletado entre 2012 e 2019, com foco em corredores de médias e longas distâncias. A metodologia incluirá o pré-processamento dos dados, a divisão em conjuntos de treinamento e testes, e a criação de um modelo preditivo baseado em técnicas de aprendizado supervisionado. O modelo será validado utilizando métricas como acurácia, precisão, *Recall* e *F2-score*. Espera-se que o modelo preditivo seja capaz de identificar corredores em risco de lesões com boa precisão e oferecer recomendações personalizadas baseadas na carga de treinamento e intensidade.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Introdução à Corrida e suas Métricas

A corrida de rua praticada de maneira recreativa ou de maneira competitiva é um dos exercícios físicos mais antigos e acessíveis. Sua simplicidade, aliada à ausência de necessidade de grandes investimentos ou infraestrutura específica tem contribuído para o crescimento significativo de sua popularidade nos últimos anos. Esse aumento é impulsionado também pela busca por um estilo de vida mais saudável e pelo crescimento do número de eventos esportivos ao redor do mundo. Além disso, o surgimento e consolidação de aplicativos de monitoramento esportivo, como Strava[1] e Garmin Connect[2] por exemplo, contribuiu para o aumento da adesão, pois permite que corredores acompanhem seu progresso e interajam com outros praticantes.

De acordo com (Toledo, 2024) um relatório do Strava apontou que a corrida foi o esporte mais praticado globalmente em 2024. No Brasil, o número de clubes de corrida cresceu 109%, quase o dobro da média global de 59%. Ademais, a maior difusão de informação sobre treinamento, prevenção de lesões e nutrição tem incentivado mais pessoas a adotarem a corrida como uma atividade regular.

Com a evolução da tecnologia e da ciência do esporte, diversas métricas foram desenvolvidas para monitorar e aprimorar o desempenho dos corredores. Entre as principais métricas utilizadas na corrida, destacam-se a distância percorrida, o ritmo (*pace*), a frequência cardíaca, a cadência e o tempo de contato com o solo. O ritmo é medido em minutos por quilômetro e é uma das principais referências para avaliar a intensidade do treino. A frequência cardíaca, expressa em batimentos por minuto (bpm), serve como parâmetro para o controle da carga de esforço. A cadência, que corresponde ao número de passos por minuto, e o tempo de contato com o solo são métricas biomecânicas que impactam a eficiência da corrida.

Entre os parâmetros possíveis de serem monitorados mencionados acima, a frequência cardíaca merece um destaque especial. Além de entender o esforço momentâneo, utilizando-se da mesma é possível também definir as chamadas zonas de treinamento, que são faixas de intensidade baseadas na frequência cardíaca máxima (FCM) de um indivíduo, utilizadas para otimizar os objetivos de treino. Elas variam desde a recuperação (50-60% da FCM) até o esforço máximo (90-100% da FCM), sendo aplicadas para melhorar resistência, velocidade, força explosiva e capacidade aeróbica (Fleck; Kraemer, 2014).

[1] STRAVA. *Strava*. Disponível em: https://strava.com/

[2] GARMIN. *Garmin Connect*. Disponível em: https://connect.garmin.com/.

2.2 Modelos de treinamentos para corrida

Visando o ganho de performance minimizando o risco de lesões, os corredores utilizam diferentes tipos de treinos, cada um com um objetivo específico. A combinação adequada desses treinos permite um desenvolvimento equilibrado das capacidades aeróbicas, anaeróbicas e musculoesqueléticas, resultando em uma evolução progressiva e sustentável. A seguir, são apresentados os principais tipos de treinamento na corrida.

2.2.1 Treino de Rodagem

O treino de rodagem consiste em corridas de baixa a moderada intensidade, geralmente realizadas em ritmos confortáveis, nos quais o corredor consegue manter uma conversa sem grande esforço. Esse tipo de treino tem como principais objetivos:

- Desenvolver a base aeróbica, essencial para qualquer corredor, independentemente do nível de experiência;
- Melhorar a resistência cardiovascular e muscular;
- Ajudar na recuperação entre treinos mais intensos.

2.2.2 Treino Intervalado

O treino intervalado caracteriza-se pela alternância entre períodos de alta intensidade e momentos de recuperação ativa (trote ou caminhada). Os principais benefícios desse treino incluem:

- Melhora da velocidade e potência aeróbica máxima (VO_2 máximo);
- Aumento da eficiência cardiorrespiratória;
- Desenvolvimento da capacidade anaeróbica, permitindo sustentar esforços intensos por mais tempo;
- Maior queima calórica e adaptação neuromuscular.

2.2.3 Treino de Ritmo (Tempo *Run*)

O treino de ritmo, também chamado de "tempo *run*", é realizado em uma intensidade próxima ao limiar anaeróbico, que corresponde ao maior ritmo que o atleta consegue sustentar por um período prolongado sem acumular fadiga excessiva. Seus objetivos incluem:

- Melhorar a capacidade de manter ritmos elevados por longos períodos;

- Aumentar a tolerância ao acúmulo de lactato no sangue;
- Desenvolver a resistência mental e física para competições.

2.2.4 Treino Longo

O treino longo é um tipo de treinamento considerado fundamental para provas de maiores distâncias onde se é exigida resistência por parte dos atletas. O treinamento envolve corridas de longa duração e baixa intensidade, onde o foco principal é a adaptação fisiológica e mental ao esforço prolongado. Seus benefícios incluem:

- Aprimoramento da capacidade aeróbica e da resistência muscular;
- Aumento da eficiência na utilização de gorduras como fonte de energia, o que ajuda na economia de glicogênio;
- Adaptação psicológica e estabilidade emocional para enfrentar a fadiga nas fases finais das provas longas.

Cada tipo de treino pode ser categorizado com base na zona de frequência cardíaca em que ocorre, o que reflete a intensidade e os benefícios fisiológicos esperados. O treino de rodagem, realizado de forma contínua e sem grande exigência fisiológica, ocorre geralmente entre 60-75% da frequência cardíaca máxima (FCmáx), favorecendo o desenvolvimento aeróbico e a recuperação muscular. O treino longo, por sua vez, também se mantém nessa faixa, mas com um foco maior na resistência e na capacidade de sustentar o esforço por períodos prolongados. O treino de ritmo, voltado para a melhora do limiar anaeróbico, ocorre entre 80-90% da FCmáx, desenvolvendo a capacidade de manter altas intensidades sem acúmulo excessivo de fadiga. Já o treino intervalado, que busca estimular adaptações neuromusculares e aumento do VO2 máximo, consiste em esforços intensos acima de 90% da FCmáx, intercalados com períodos curtos de recuperação. O controle dessas zonas é essencial para garantir a efetividade do treinamento e minimizar o risco de lesões (Seiler, 2010).

2.3 Intensidade do Treinamento e o Risco de Lesões

A relação entre a intensidade do treinamento e o risco de lesões tem sido amplamente estudada na literatura esportiva. Treinos de alta intensidade, como os intervalados e os de ritmo (tempo *run*), impõem uma carga significativa ao sistema musculoesquelético, aumentando o risco de microlesões, inflamações e fadiga excessiva. Segundo (Seiler, 2010), o estresse fisiológico causado por treinos realizados acima de 90% da frequência cardíaca máxima (FCmáx) pode comprometer a recuperação muscular e elevar a probabilidade de lesões por sobrecarga.

Dessa forma, um planejamento adequado do treinamento, respeitando a distribuição entre treinos leves, moderados e intensos, é fundamental para otimizar o desempenho e minimizar riscos. Estratégias como a periodização e o controle da carga de treinamento ajudam a equilibrar os estímulos e a reduzir a incidência de lesões (Gabbett, 2016). Além disso, fatores como descanso adequado, nutrição e fortalecimento muscular também desempenham um papel essencial na prevenção de lesões associadas à alta intensidade do treinamento (Soligard *et al.*, 2016).

2.3.1 Parâmetros de Carga e Fadiga Acumulada

A quantificação da carga de treinamento é essencial para o monitoramento do atleta. O *Acute Chronic Workload Ratio* (ACWR) tem emergido como um importante indicador de risco de lesão, representando a relação entre a carga aguda (normalmente 1 semana) e a carga crônica (4 semanas) (Blanch; Gabbett, 2016). Valores acima de 1,5 são consistentemente associados com aumento significativo do risco de lesões, enquanto ratios entre 0,8-1,3 parecem oferecer proteção contra lesões.

A fadiga acumulada manifesta-se através de múltiplos indicadores, incluindo diminuição do desempenho, alterações na variabilidade da frequência cardíaca, aumento da percepção subjetiva de esforço para mesma carga, e distúrbios do sono (Kellmann *et al.*, 2018). A monitorização regular destes parâmetros permite intervenções precoces antes que a fadiga excessiva resulte em lesão.

2.3.2 Principais Tipos de Lesão em Corredores

As lesões por sobrecarga predominam entre corredores, com incidências variando de 19% a 79% anualmente (Gent *et al.*, 2007). As principais categorias incluem:

- Lesões Musculares: Estiramentos e distensões da cadeia posterior, particularmente dos isquiotibiais e gastrocnêmios
- Tendinopatias: Tendinopatia do Aquiles, tendinopatia patelar e síndrome da banda iliotibial
- Lesões Ósseas: Fraturas por estresse da tíbia, metatarsos e fíbula
- Síndromes Dolorosas: Síndrome da dor patelofemoral, fascite plantar e síndrome do estresse tibial medial

Estas lesões frequentemente resultam de erros de treinamento, incluindo aumentos abruptos de volume ou intensidade, recuperação inadequada e défices biomecânicos não corrigidos (Buist *et al.*, 2010).

2.3.3 Fatores de Risco e Estratégias Preventivas

Os fatores de risco para lesões em corredores podem ser categorizados em intrínsecos (idade, sexo, histórico prévio de lesão, défices de flexibilidade ou força) e extrínsecos (volume e intensidade de treino, calçado, superfície de treino) (Saragiotto *et al.*, 2014).

Estratégias preventivas eficazes incluem:

- Progressão Gradual: Limitar aumentos semanais de volume a não mais que 10%
- Periodização Adequada: Inclusão de ciclos de recuperação e treino de força
- Correção Biomecânica: Abordagem de déficits de mobilidade e força
- Monitorização Contínua: Uso de tecnologias para monitoramento de carga e recuperação

A integração destas estratégias num abordagem abrangente permite otimizar o desempenho enquanto minimiza o risco de interrupções por lesão (Bahr, 2012).

2.4 Abordagens de Aprendizado de Máquina e *Deep Learning* para Modelagem Preditiva de Lesões em Corredores

A previsão de lesões em corredores é um problema complexo que pode ser abordado por diferentes técnicas de Inteligência Artificial. Com base nos dados coletados, métodos de aprendizado de máquina e aprendizado profundo podem ser utilizados para identificar padrões e prever o risco de lesão com base na carga de treinamento, intensidade, frequência cardíaca, e outras variáveis.

2.4.1 Aprendizado Supervisionado e Modelos de Classificação

O aprendizado supervisionado é uma abordagem amplamente utilizada quando há um conjunto de dados rotulado, ou seja, quando as ocorrências de lesões são conhecidas e podem ser associadas a padrões prévios. Para o problema de prevenção de lesão em corredores, modelos de classificação como Regressão Logística e *Random Forest* podem ser empregados (Luger, 2017). Esses algoritmos podem analisar características como volume, intensidade dos treinos e variabilidade dos mesmos visando estimar a probabilidade de ocorrência de lesões.

A Regressão Logística e o *Random Forest* são duas técnicas amplamente utilizadas em aprendizado de máquina para tarefas de classificação e regressão. A Regressão Logística é uma técnica de classificação que modela a probabilidade de um evento binário, como a ocorrência de uma lesão, por meio de uma combinação linear de variáveis preditoras, mapeada pela função sigmoide:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

onde $\mathbf{w}$ são os pesos e b é o viés. O modelo é treinado para otimizar os parâmetros $\mathbf{w}$ e b minimizando a entropia cruzada, ajustada para dados desbalanceados com pesos de classe que priorizam a classe minoritária (lesões) (Hastie; Tibshirani; Friedman, 2009a). Sua principal vantagem é a interpretabilidade, permitindo analisar o impacto de características na probabilidade de lesão (Carey *et al.*, 2017). No contexto de esportes, a Regressão Logística é usada para modelar relações lineares entre métricas de treinamento e riscos de lesão, fornecendo previsões estáveis que complementam métodos mais complexos, como redes neurais (Claudino *et al.*, 2019).

O *Random Forest* é um método que combina previsões de múltiplas árvores de decisão para melhorar a precisão e reduzir o *overfitting*:

$$\hat{y}_{\text{RF}} = \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{x})$$

onde $h_t(\mathbf{x})$ é a previsão da árvore t e T é o número de árvores. O *Random Forest* complementa a LSTM ao processar dados tabulares, oferecendo previsões robustas e reduzindo a variância em cenários desbalanceados (Claudino *et al.*, 2019).

2.4.2 Boosting

O *boosting* é uma técnica que constrói modelos de forma sequencial, onde cada novo modelo tenta corrigir os erros cometidos por seus predecessores. Dessa forma, modelos fracos (isto é, preditores com desempenho ligeiramente superior ao acaso) são combinados para formar um modelo forte. Diferentemente do *bagging*, o *boosting* atribui pesos maiores às instâncias mais difíceis de classificar, enfatizando exemplos mal classificados ao longo do processo de treinamento (Freund; Schapire, 1997).

Entre os algoritmos mais conhecidos de *boosting* destacam-se:

2.4.2.1 *AdaBoost*

O *AdaBoost*, proposto por Freund e Schapire (Freund; Schapire, 1997), ajusta iterativamente pesos às amostras do conjunto de dados, de modo que as instâncias mais difíceis de classificar recebam maior atenção nos modelos subsequentes.

2.4.2.2 *Gradient Boosting*

O *Gradient Boosting* (Friedman, 2001) constrói modelos aditivos sequenciais, minimizando uma função de perda por gradiente. Cada novo modelo é treinado para corrigir os erros residuais do modelo anterior, resultando em uma combinação eficiente de modelos fracos em um modelo forte.

2.4.2.3 *XGBoost*

O *XGBoost* (*Extreme Gradient Boosting*) (Chen; Guestrin, 2016) é uma implementação otimizada de *Gradient Boosting* que incorpora regularização para evitar sobreajuste, paralelização de processamento e suporte a dados esparsos. Essa técnica tem se destacado em competições de aprendizado de máquina devido à sua eficiência computacional e alta acurácia. Além disso, permite ajustes finos de hiperparâmetros, como taxa de aprendizado, profundidade das árvores e número de estimadores, tornando-o extremamente flexível para diferentes tipos de dados e problemas de predição.

2.4.3 Redes Neurais Recorrentes e Arquitetura LSTM: Modelagem de Dependências Temporais em Séries de Treinamento

As Redes Neurais Recorrentes (RNNs) representam uma classe de arquiteturas de *Deep Learning* especificamente projetadas para processar dados sequenciais, tornando-se particularmente adequadas para a modelagem de séries temporais de treinamento esportivo. No entanto, as RNNs tradicionais sofrem do problema de *vanishing gradients*, que limita sua capacidade de aprender dependências de longo prazo em sequências extensas. Para superar esta limitação, as *Long Short-Term Memory (LSTM) networks* foram desenvolvidas com um mecanismo de portões (*gating mechanism*) que permite o controle preciso do fluxo de informação através do tempo (Hochreiter; Schmidhuber, 1997).

A arquitetura LSTM é baseada em uma célula de memória $\mathbf{C}_t$ e três portões não-lineares que regulam as operações de leitura, escrita e reset da memória. As equações que governam a dinâmica temporal em cada passo t são:

$$\text{Portão de Esquecimento (}\textit{Forget Gate}\text{):} \quad \mathbf{f}_t = \sigma(\mathbf{W}_f \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f) \tag{2.1}$$

$$\text{Portão de Entrada (}\textit{Input Gate}\text{):} \quad \mathbf{i}_t = \sigma(\mathbf{W}_i \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i) \tag{2.2}$$

$$\text{Célula de Memória Candidata:} \quad \tilde{\mathbf{C}}_t = \tanh(\mathbf{W}_C \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_C) \tag{2.3}$$

$$\text{Atualização da Memória:} \quad \mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{C}}_t \tag{2.4}$$

$$\text{Portão de Saída (}\textit{Output Gate}\text{):} \quad \mathbf{o}_t = \sigma(\mathbf{W}_o \cdot [\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o) \tag{2.5}$$

$$\text{Estado Oculto:} \quad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{C}_t) \tag{2.6}$$

onde σ representa a função sigmoide, $\odot$ denota o produto de Hadamard (*element-wise*), $\mathbf{x}_t$ é o vetor de entrada no tempo t, $\mathbf{h}_t$ é o estado oculto, e $\mathbf{W}$ e $\mathbf{b}$ são parâmetros aprendíveis.

No contexto da previsão de lesões em corredores, a LSTM demonstra capacidades excepcionais para: (1) Modelagem de Longo Prazo: capturar padrões de acumulação de fadiga e adaptação fisiológica que se desenvolvem ao longo de várias semanas de treinamento; (2) Identificação de Padrões Temporais Complexos: reconhecer sequências

críticas de carga alta seguida por recuperação insuficiente que precedem eventos lesivos; (3) Processamento de Dados Multivariados: integrar informações heterogêneas incluindo volume de treino, intensidade, frequência cardíaca, e métricas de recuperação em uma representação temporal coerente; e (4) Generalização para Comportamentos Não Vistos: inferir padrões de risco mesmo para sequências de treinamento não observadas durante o treino, crucial para a aplicação em novos atletas (Graves, 2012).

A capacidade da LSTM de memorizar estados relevantes através de longos intervalos temporais é particularmente valiosa para a previsão de lesões por sobrecarga, onde o efeito cumulativo de cargas de treinamento inadequadas pode manifestar-se apenas após várias semanas de repetição. Esta característica distingue as LSTMs de modelos tradicionais de *Machine Learning* que tratam cada observação como independente, ignorando a natureza temporal inerente aos dados de treinamento esportivo (Sherstinsky, 2020).

2.4.4 *Ensemble Learning*: Conceito e Estratégias

O *ensemble learning* é uma abordagem ampla em aprendizado de máquina que tem como objetivo combinar as previsões de múltiplos modelos para produzir uma decisão mais precisa e robusta do que qualquer modelo individual, reduzindo erros de variância, viés ou ambos (Hastie; Tibshirani; Friedman, 2009b).

Enquanto técnicas como o *boosting* e suas variantes (por exemplo, o XGBoost) representam implementações específicas de *ensemble*, o conceito é mais abrangente e permite a combinação de modelos de diferentes naturezas. Essa estratégia é particularmente eficaz em problemas complexos, como a previsão de lesões em corredores, nos quais dados desbalanceados e padrões temporais desafiam a performance de um único algoritmo (Rokach, 2010; Sagi; Rokach, 2018).

No contexto do *ensemble*, cada modelo contribui com suas forças: modelos simples, como a Regressão Logística, capturam relações lineares; modelos baseados em árvores, como *Random Forest* e *XGBoost*, modelam relações não lineares; e redes neurais, como a LSTM, capturam dependências temporais (Claudino *et al.*, 2019). A combinação pode ser feita por métodos como votação ponderada:

$$\hat{y}_{\text{ensemble}} = \sum_{m=1}^{M} w_m \hat{y}_m, \quad \sum_{m=1}^{M} w_m = 1,$$

onde $\hat{y}_m$ é a previsão do modelo m e w_m é seu peso, ou por *stacking*, onde um meta-modelo (e.g., Regressão Logística) aprende a combinar as previsões:

$$\hat{y}_{\text{final}} = \sigma(\mathbf{w}^T [\hat{y}_{\text{model1}}, \hat{y}_{\text{model2}}, \dots] + b).$$

As principais estratégias de *ensemble* incluem: (i) *bagging* (Bootstrap Aggregating), que treina modelos independentes em subamostras aleatórias dos dados, reduzindo variância,

como no Random Forest (Breiman, 2001); (ii) *boosting*, que constrói modelos sequencialmente, onde cada modelo corrige os erros do anterior, como no *XGBoost* (Chen; Guestrin, 2016); e (iii) *stacking*, que usa um meta-modelo para integrar previsões de modelos heterogêneos, como no pipeline (Wolpert, 1992).

É fundamental compreender a relação hierárquica entre *Ensemble Learning* e *XGBoost* para contextualizar adequadamente sua aplicação neste trabalho. O *Ensemble Learning* constitui um paradigma amplo que engloba qualquer metodologia que combine múltiplos modelos para melhorar o desempenho preditivo. Dentro deste paradigma, o *XGBoost* representa uma implementação específica e altamente otimizada do algoritmo de *Gradient Boosting*, que por sua vez é uma categoria particular de métodos de ensemble.

A relação pode ser expressa formalmente através da seguinte hierarquia:

$$\textit{Ensemble Learning} \supset \textit{Boosting} \supset \textit{Gradient Boosting} \supset \textit{XGBoost} \tag{2.7}$$

Neste trabalho, adotou-se uma estratégia de ensemble heterogêneo onde o XGBoost atua como um dos componentes especializados. A arquitetura geral do ensemble pode ser representada por:

$$\hat{y}_{\text{ensemble}} = \sum_{j=1}^{M} w_j \cdot f_j(\mathbf{x}) \tag{2.8}$$

onde f_j representa cada modelo base (incluindo o *XGBoost*) e w_j seus respectivos pesos.

O XGBoost contribui para o ensemble através de sua capacidade especializada em:

- Capturar relações não-lineares complexas através de seu mecanismo de *boosting* sequencial
- Lidar eficientemente com *missing values* e dados heterogêneos
- Oferecer excelente desempenho em dados tabulares através de sua implementação otimizada

Sua função objetivo regularizada, dada por:

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{2.9}$$

onde $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2$, fornece forte controle contra *overfitting*, complementando outros modelos do *ensemble*.

A integração do *XGBoost* dentro de um ensemble maior justifica-se pela diversidade de modelos, onde:

$$\text{Error}_{\text{ensemble}} \leq \overline{\text{Error}} - \overline{\text{Diversity}} \tag{2.10}$$

Esta abordagem híbrida permite que o ensemble aproveite as melhores características de cada algoritmo, resultando em um sistema preditivo mais preciso e confiável para a tarefa complexa de previsão de lesões em corredores.

2.4.5 Pré-processamento de Dados para Modelagem Preditiva

O pré-processamento de dados constitui a etapa fundamental de preparação e transformação dos dados brutos para alimentar algoritmos de aprendizado de máquina. Em problemas de predição de lesões esportivas, esta fase é particularmente crítica devido às características específicas dos dados de treinamento atlético (Gent *et al.*, 2007).

A imputação de valores ausentes segue uma estratégia hierárquica que prioriza a individualidade biológica. Inicialmente, valores ausentes são preenchidos com a mediana específica de cada atleta, preservando assim suas características fisiológicas únicas. Apenas quando insuficientes dados do atleta estão disponíveis recorre-se à mediana global, garantindo robustez estatística (Little; Rubin, 2019):

$$x_{\text{imputado}} = \begin{cases} \text{mediana}_{\text{atleta}}(x) & \text{se } n_{\text{atleta}} \geq 3 \\ \text{mediana}_{\text{global}}(x) & \text{caso contrário} \end{cases} \tag{2.11}$$

A normalização robusta é aplicada utilizando técnicas resistentes a *outliers*, como o *RobustScaler*, que baseia-se na mediana e no intervalo interquartil ao invés da média e desvio padrão tradicionais (Huber; Ronchetti, 2011):

$$x_{\text{scaled}} = \frac{x - \text{mediana}(x)}{\text{IQR}(x)} \tag{2.12}$$

Esta abordagem é especialmente adequada para dados fisiológicos onde valores extremos podem ocorrer naturalmente sem necessariamente representar erro de medição.

O balanceamento de classes aborda o desbalanceamento extremo característico desses problemas (tipicamente <5% de casos positivos). Uma estratégia híbrida combina *undersampling* da classe majoritária com *oversampling* sintético da classe minoritária através do BorderlineSMOTE, que gera exemplos sintéticos nas regiões de fronteira de decisão, onde a classificação é mais incerta (Han; Wang; Mao, 2005; Chawla *et al.*, 2002).

A validação cruzada por grupo (*Group K-Fold*) é essencial para evitar vazamento de informação, garantindo que todos os dados de um mesmo atleta permaneçam no mesmo *fold* durante a validação. Esta abordagem fornece estimativas realistas de performance para novos atletas não vistos durante o treinamento (Arlot; Celisse, 2010).

2.4.6 Engenharia de *Features* para Dados Temporais Esportivos

A engenharia de *features* transforma dados brutos de treinamento em variáveis temporalmente significativas que capturam padrões fisiológicos e biomecânicos relevantes para a predição de lesões. No contexto de corredores, esta etapa é crucial para representar adequadamente a dinâmica de carga, recuperação e adaptação (Gabbett, 2016).

Dentre as *features* temporais desenvolvidas, destacam-se as métricas de carga relativa, como o *Acute:Chronic Workload Ratio* (ACWR), que quantifica a relação entre a carga aguda (7 dias) e crônica (28 dias), servindo como indicador de risco quando excede valores críticos estabelecidos na literatura esportiva (Blanch; Gabbett, 2016; Hulin *et al.*, 2014).

Médias móveis de diferentes janelas temporais (3, 7, 28 dias) capturam tendências de curto, médio e longo prazo no volume e intensidade do treinamento, enquanto medidas de variabilidade (desvio padrão, coeficiente de variação) indicam a consistência das cargas aplicadas (Impellizzeri *et al.*, 2004).

Features de mudança percentual e taxa de variação permitem identificar mudanças abruptas na carga de treinamento, frequentemente associados ao aumento do risco de lesão (Colby *et al.*, 2017). Variáveis de acumulação progressiva detectam períodos de aumento sustentado de carga, enquanto indicadores de desbalanço entre esforço e recuperação capturam situações onde a carga aplicada excede consistentemente a capacidade de recuperação (Soligard *et al.*, 2016).

A seleção de *features* emprega uma abordagem híbrida que combina importância baseada em *Random Forest* com informação mútua, priorizando variáveis com significância estatística e relevância clínica (Guyon; Elisseeff, 2003). *Features* críticas de domínio como ACWR e razão de recuperação são preservadas independentemente de seus scores estatísticos (Bourdon *et al.*, 2017).

2.4.7 Métricas de Avaliação para Problemas Desbalanceados

A avaliação de modelos em problemas de predição de lesões requer métricas especializadas que contemplem a natureza desbalanceada dos dados. Métricas convencionais como acurácia tornam-se enganosas, podendo atingir valores altos mesmo quando o modelo falha completamente em detectar casos positivos (He; Garcia, 2009).

O *Recall* (Sensibilidade) emerge como métrica principal, representando a capacidade do modelo de identificar corretamente os verdadeiros positivos (Saito; Rehmsmeier, 2015):

$$\text{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{2.13}$$

onde VP são os verdadeiros positivos e FN os falsos negativos. No contexto de prevenção de lesões, falsos negativos possuem consequências particularmente severas.

O F_2*-Score* amplia esta priorização, atribuindo maior peso ao *Recall* em relação à Precisão (Chinchor, 1992). A fórmula geral do F_β-Score é dada por:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precisão} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precisão}) + \text{Recall}} \tag{2.14}$$

No caso de $\beta = 2$, tem-se:

$$F_2 = 5 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{4 \cdot \text{Precisão} + \text{Recall}} \tag{2.15}$$

A Área Sob a Curva *Precision-Recall* (AUC-PR) mostra-se mais informativa que a AUC-ROC tradicional para dados desbalanceados, pois focaliza explicitamente no desempenho da classe minoritária (Davis; Goadrich, 2006).

A *Balanced Accuracy* fornece uma visão balanceada do desempenho, calculando a média aritmética da sensibilidade e especificidade (Brodersen *et al.*, 2010):

$$\text{Balanced Accuracy} = \frac{\text{Sensibilidade} + \text{Especificidade}}{2} \tag{2.16}$$

A otimização de limiares é realizada para maximizar o *Recall* enquanto mantém uma precisão mínima clinicamente aceitável, reconhecendo que no contexto de prevenção de lesões, a falha em detectar um caso positivo (falso negativo) é significativamente mais crítica que um falso alarme (falso positivo) (Provost; Fawcett; Kohavi, 2001).

Estas métricas, quando consideradas em conjunto, fornecem uma avaliação abrangente e clinicamente relevante do desempenho preditivo dos modelos desenvolvidos.

3 METODOLOGIA

3.1 Descrição dos Dados

Este estudo utilizou dois conjuntos de dados públicos disponibilizados por (Lovdal; Hartigh; Azzopardi, 2020) et al., coletados de uma equipe de corrida de alto rendimento dos Países Baixos entre 2012 e 2019. A amostra inclui 74 atletas de ambos os sexos, especializados em provas de média e longa distância (800 metros à maratona), com rotinas de treinamento predominantemente de resistência. A consistência metodológica foi garantida pelo acompanhamento de um único treinador durante todo o período de coleta. O estudo foi aprovado por comitê de ética e conduzido conforme a Declaração de Helsinque.

Os conjuntos de dados utilizados foram:

- **day_approach_maskedID_timeseries.csv**: Dados diários com 42.766 entradas e 73 características, incluindo a variável alvo *injury* e identificador temporal *Date*. Contém métricas de treinamento (número de sessões, quilometragem total e por zonas de intensidade), horas de treinamento alternativo, exercícios de força e métricas de percepção (esforço, sucesso no treinamento, recuperação).
- **week_approach_maskedID_timeseries.csv**: Dados semanais com 42.798 entradas e 72 características, também incluindo *injury* e *Date*. Apresenta features semelhantes ao conjunto diário, porém agregadas em nível semanal e em diferentes janelas temporais, incorporando medidas relativas de quilometragem, intensidade, treinos de força, recuperação e percepções subjetivas.

Ambos os conjuntos incluem *Athlete ID* para identificação individual e a variável *injury* (1 para lesão, 0 para não lesão).

3.2 Preparação e Pré-processamento de Dados

O processo iniciou com a inspeção e limpeza dos dados, removendo valores duplicados e corrigindo inconsistências. Realizou-se imputação de valores ausentes utilizando a mediana de cada variável, preservando a distribuição original dos dados. As fontes de dados foram harmonizadas através de agregação temporal, criando uma estrutura unificada para análise.

3.3 Engenharia de *Features*

Desenvolveu-se *features* derivadas para capturar padrões relevantes, incluindo:

- Métricas de tendência central (média, mediana) e dispersão (desvio padrão, amplitude) em janelas temporais variadas
- Variações percentuais para identificar mudanças abruptas
- Indicadores de estabilidade para avaliar consistência comportamental

A seleção final considerou correlação com a variável alvo e redundância entre preditores, priorizando *features* estatisticamente significativas.

3.4 Modelagem Preditiva

Implementou-se uma abordagem comparativa com quatro modelos individuais:

- Regressão Logística: Configurada com regularização L2 para evitar *overfitting*
- *Random Forest*: Parametrização padrão focada em generalização
- *XGBoost*: Com regularização avançada e tratamento eficiente de dados
- LSTM: Projetada para capturar dependências temporais nas sequências

Adicionalmente, desenvolveu-se um modelo *Ensemble* baseado em votação majoritária ponderada pelo desempenho individual de cada modelo base.

3.5 Validação e Avaliação

Considerando a natureza desbalanceada dos dados (aprox. 1,34% de exemplos positivos), adotou-se validação cruzada estratificada em 3 *folds*, preservando a distribuição original das classes em cada divisão.

A avaliação priorizou métricas adequadas para problemas desbalanceados:

- ***Recall***: Sensibilidade na detecção de casos positivos
- ***F2-Score***: Ponderação que valoriza recall sobre precisão
- **AUC-ROC** e ***Balanced Accuracy***: Medidas abrangentes de desempenho

Manteve-se o limiar de classificação padrão (0,5) para permitir comparação direta entre modelos.

3.6 Fluxo de Implementação

O processo seguiu um fluxo sequencial documentado na Figura 1, abrangendo desde o pré-processamento até a avaliação final, com documentação completa para garantir reprodutibilidade.

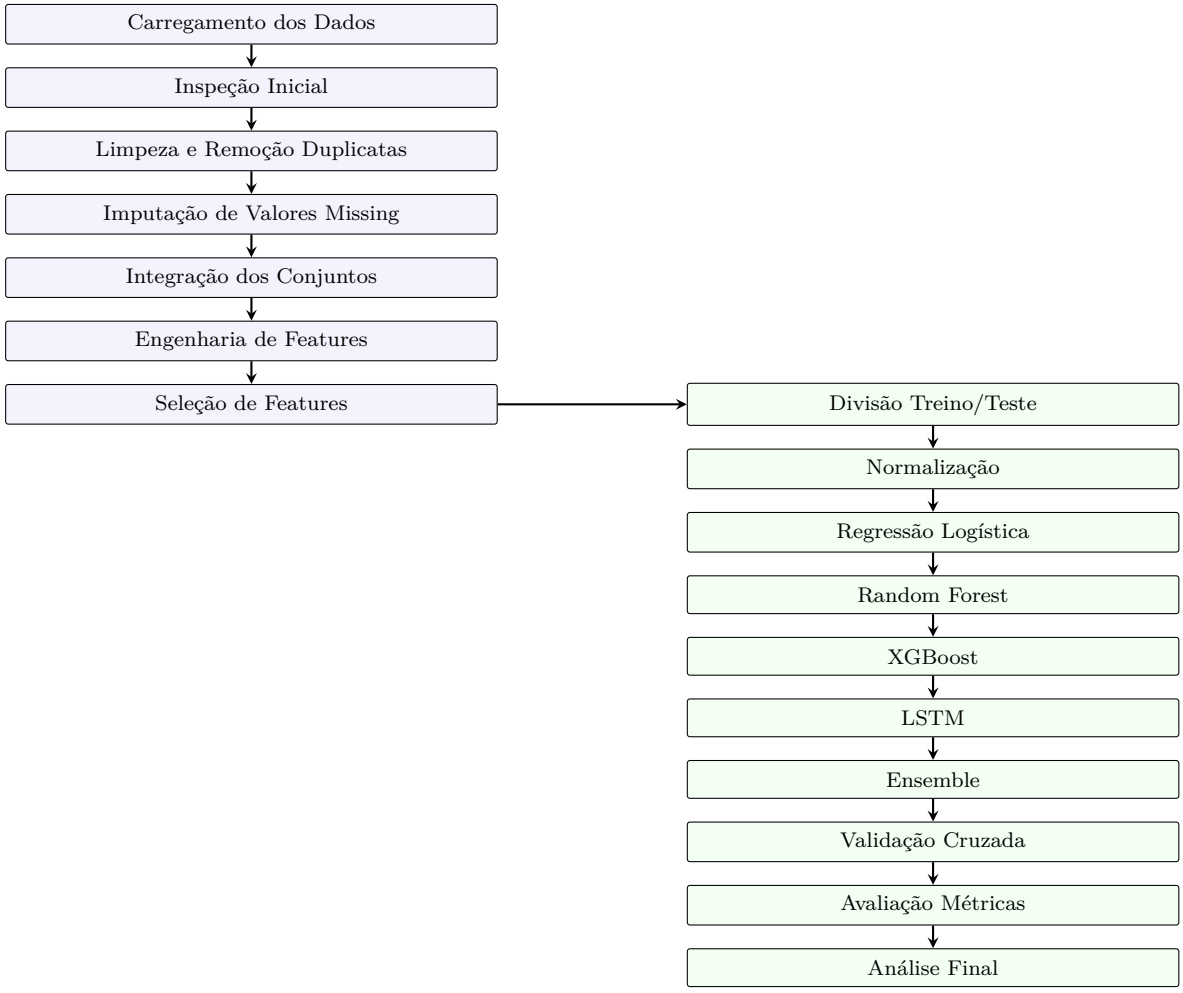

Figura 1 – Fluxo completo do processo de desenvolvimento: desde o pré-processamento até a avaliação final

4 AVALIAÇÃO EXPERIMENTAL

4.1 Análise da Importância das *Features*

A análise da importância das *features* é crucial para compreender quais variáveis têm maior poder preditivo no modelo *Ensemble* desenvolvido. Esta análise foi realizada utilizando duas abordagens complementares: importância baseada no *Random Forest* e scores de informação mútua, combinadas em um score final ponderado.

4.1.1 Metodologia de Análise

A importância das features foi calculada através de:

- Importância do Random Forest: Medida baseada na redução média de impureza (Gini) proporcionada por cada *feature* em todas as árvores do modelo
- Informação Mútua (MI): Medida não-paramétrica que quantifica a dependência estatística entre cada *feature* e a variável *target*
- *Score* Combinado: Média ponderada das duas métricas anteriores, proporcionando uma visão abrangente da relevância de cada feature

4.1.2 *Features* Mais Relevantes

A Tabela 1 apresenta as 20 *features* mais importantes identificadas pela análise.

4.1.3 Interpretação Detalhada das *Features* Mais Importantes

As *features* mais relevantes podem ser categorizadas em quatro grupos principais:

1. Métricas de Volume e Carga de Treinamento:

- **total km_ma3** e **total km_ma7**: Médias móveis de 3 e 7 dias da quilometragem total percorrida, indicando tendências recentes de volume de treino
- **km Z3-4_ma3** e **km Z3-4_ma7**: Médias móveis de quilometragem em zonas de intensidade 3-4 (corrida no limiar anaeróbico ou ligeiramente acima)
- **load_accumulation_28d**: Acumulação de carga em 28 dias, representando o volume total de treino recente
- **km sprinting**: Volume total de quilômetros percorridos em *sprints* (máxima intensidade)

2. Indicadores de Intensidade e Esforço Percebido:

Feature	MI Score	RF Importance	Combined Score
total km_ma3	0.00366	0.00333	0.00356
km Z3-4_ma7	0.00309	0.00407	0.00339
load_accumulation_28d	0.00176	0.00513	0.00277
max exertion_std_all_change_7d	0.00096	0.00892	0.00335
hours alternative_mean_all	0.00367	0.00253	0.00333
total km Z3-Z4-Z5-T1-T2_mean_all_-overtraining	0.00202	0.00622	0.00328
km sprinting	0.00307	0.00352	0.00320
nr. strength trainings	0.00255	0.00456	0.00316
km Z3-4_ma3	0.00285	0.00373	0.00311
hours alternative	0.00348	0.00176	0.00296
strength training_ma3	0.00249	0.00389	0.00291
total km_ma7	0.00243	0.00361	0.00278
strength training_ma7	0.00219	0.00447	0.00287
max recovery_std_all	0.00112	0.00558	0.00246
total kms_std_all	0.00000	0.00574	0.00172
min exertion_std_all	0.00115	0.00434	0.00210
avg recovery_std_all	0.00056	0.00727	0.00258
min recovery_std_all	0.00048	0.00808	0.00276
nr. sessions_std_all	0.00000	0.00574	0.00172
total km Z5-T1-T2_std_all_change_7d	0.00097	0.00592	0.00246

Tabela 1 – Top 20 *Features* por Importância Combinada

- **max exertion_std_all_change_7d**: Variação no esforço percebido máximo (avaliação subjetiva do atleta sobre a dificuldade do treino), indicando flutuações significativas na percepção de intensidade
- **total km Z3-Z4-Z5-T1-T2**: Quilometragem total em zonas de alta intensidade (Z3 ou superior), incluindo treinos no limiar anaeróbico e intervalados

3. Métricas de Recuperação e Treino Complementar:

- **hours alternative_mean_all**: Média de horas dedicadas ao treino alternativo (*cross-training*), importante para recuperação ativa e prevenção de lesões
- **nr. strength trainings**: Número de sessões de treino de força realizadas, crucial para desenvolvimento de resistência muscular e prevenção de lesões
- **strength training_ma3** e **strength training_ma7**: Médias móveis de treino de força, indicando consistência no trabalho de fortalecimento

4. Variáveis de Percepção e Recuperação:

- **max recovery_std_all**: Variabilidade na percepção máxima de recuperação (quão descansado o atleta se sente antes das sessões)

- **min exertion_std_all**: Variabilidade na percepção mínima de esforço (sessões percebidas como menos intensas)
- **avg recovery_std_all**: Variabilidade na percepção média de recuperação ao longo do tempo

4.1.4 Implicações Práticas para o Treinamento

A análise demonstra que o modelo atribui maior importância a:

- **Padrões temporais consistentes**: *Features* com médias móveis (ma3, ma7) são consistentemente mais importantes que medidas pontuais, destacando a relevância de tendências e consistência no treino
- **Balanceamento carga-recuperação**: Tanto métricas de carga de treino quanto de recuperação aparecem entre as mais relevantes, enfatizando a importância deste equilíbrio
- **Percepção subjetiva do atleta**: Variáveis de esforço percebido e recuperação mostraram-se altamente preditivas, validando a importância da autoavaliação no monitoramento atlético
- **Variabilidade controlada**: *Features* que capturam variação (std_all) são importantes, sugerindo que tanto consistência quanto mudanças controladas são fatores relevantes

4.1.5 Recomendações Baseadas na Análise

Com base na importância das *features* identificadas, recomenda-se:

1. **Monitorar tendências**: Dar maior atenção às médias móveis de 3 e 7 dias em vez de valores diários isolados
2. **Manter consistência**: Desenvolver padrões consistentes de treino, especialmente em volume e intensidade
3. **Valorizar recuperação**: Incluir treino alternativo e monitorar ativamente a percepção de recuperação
4. **Gerenciar intensidade**: Controlar cuidadosamente as transições entre zonas de intensidade, especialmente para treinos de alta intensidade

Estes *insights* não apenas validam a abordagem de engenharia de *features* temporal implementada, mas também fornecem direções valiosas para intervenções práticas no planejamento e periodização do treino de atletas de corrida.

4.2 Comparativo entre Modelos Individuais e *Ensemble*

Esta seção apresenta uma análise comparativa do desempenho entre modelos individuais (Regressão Logística, *Random Forest*, *XGBoost* e LSTM) e o modelo Ensemble. A avaliação considera métricas como *Recall*, Precisão, F2-Score ($\beta = 2$), *Balanced Accuracy*, AUC-PR e AUC-ROC, complementadas por matrizes de confusão e curvas de desempenho.

4.2.1 Desempenho dos Modelos Individuais

A Tabela 2 apresenta as métricas médias obtidas por validação cruzada em 3 *folds* para cada modelo individual.

Tabela 2 – Comparativo de Desempenho dos Modelos Individuais

Modelo	**Recall**	**Precisão**	**F2-Score**	**Balanced Accuracy**	**AUC-ROC**
Regressão Logística	0,9896	0,0234	0,1069	0,4964	0,5032
Random Forest	1,0000	0,0236	0,1077	0,5000	0,6880
XGBoost	1,0000	0,0244	0,1113	0,5178	0,6207
LSTM	0,7552	0,0279	0,1208	0,5536	0,5352

4.2.1.1 Análise dos Modelos Individuais

Regressão Logística: Apresenta *Recall* elevado (0,9896), indicando poucos falsos negativos, conforme ilustrado na Figura 2. Contudo, a precisão reduzida (0,0234) e AUC-ROC próxima ao desempenho aleatório (0,5032, Figura 3) sugerem limitações na capacidade discriminativa do modelo.

Random Forest: Demonstra *Recall* perfeito (1,0000), como mostra a Figura 4, porém com precisão baixa (0,0236) e *F2-Score* moderado (0,1077). A métrica AUC-ROC (0,6880, Figura 5) apresenta valor mais expressivo em comparação aos demais modelos.

XGBoost: Apresenta comportamento similar ao *Random Forest*, com recall máximo (1,0000, Figura 6), precisão baixa (0,0244) e *F2-Score* de 0,1113. Entretanto, exibe AUC-ROC inferior (0,6207, Figura 7).

LSTM: Destaca-se pelo melhor *F2-Score* (0,1208) e *Balanced Accuracy* (0,5536), com *Recall* de 0,7552 (Figura 8) e a maior precisão entre os modelos individuais (0,0279), corroborada por AUC-ROC de 0,5352 (Figura 9).

Dentre os modelos individuais, o LSTM sobressai-se pelo equilíbrio entre *Recall* e precisão, apresentando o melhor desempenho global considerando as múltiplas métricas avaliadas.

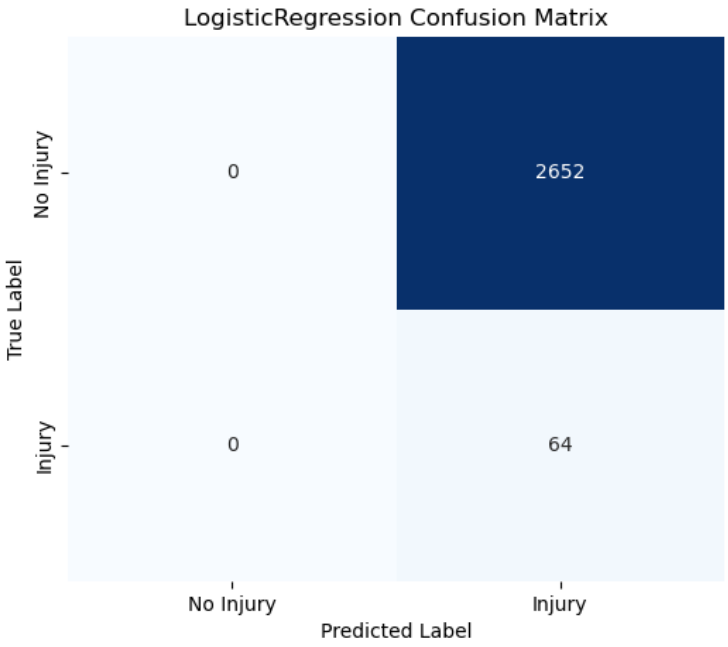

Figura 2 – Matriz de confusão da Regressão Logística.

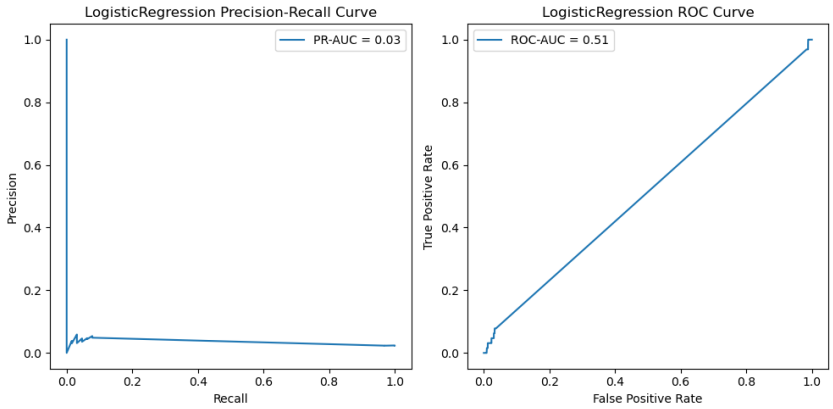

Figura 3 – Curvas PR e ROC da Regressão Logística (AUC-PR=0,0255, AUC-ROC=0,5032).

4.2.2 Análise Comparativa com o Modelo *Ensemble*

A Tabela 3 expande a análise anterior, incorporando o desempenho do modelo Ensemble.

Ensemble: O modelo Ensemble apresenta recall de 0,7031 (Figura 10), inferior aos modelos *Random Forest* e *XGBoost*, porém com a maior precisão observada (0,0444), resultando em redução significativa de falsos positivos. Seu F2-Score (0,1773) supera todos os modelos individuais, enquanto a métrica AUC-ROC (0,6926, Figura 11) demonstra competitividade, aproximando-se do valor máximo observado no *Random Forest*.

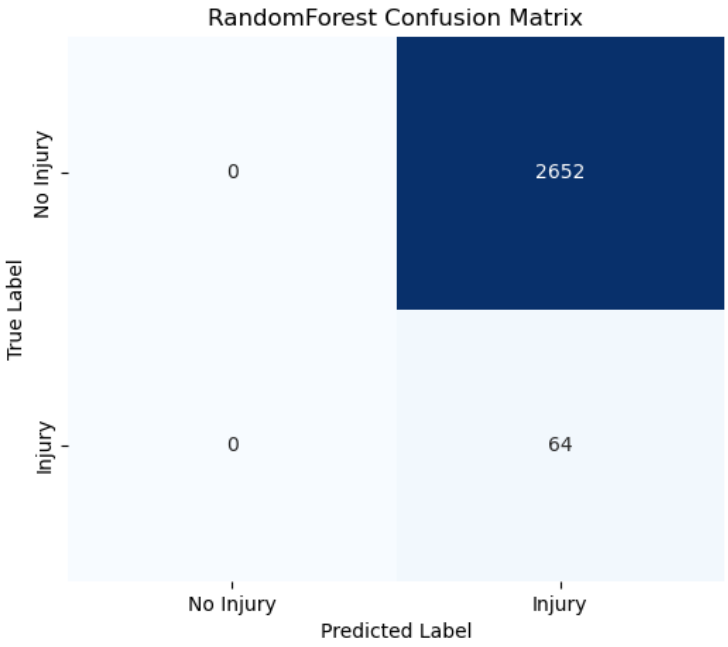

Figura 4 – Matriz de confusão do *Random Forest*.

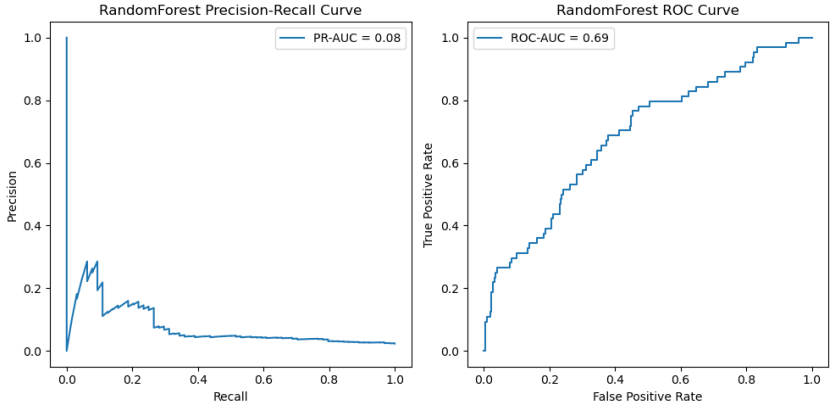

Figura 5 – Curvas PR e ROC do *Random Forest* (AUC-PR=0,0877, AUC-ROC=0,6880).

Tabela 3 – Comparativo de Desempenho: Modelos Individuais vs. *Ensemble*

Modelo	**Recall**	**Precisão**	**F2-Score**	**Balanced Accuracy**	**AUC-ROC**
Regressão Logística	0,9896	0,0234	0,1069	0,4964	0,5032
Random Forest	1,0000	0,0236	0,1077	0,5000	0,6880
XGBoost	1,0000	0,0244	0,1113	0,5178	0,6207
LSTM	0,7552	0,0279	0,1208	0,5536	0,5352
Ensemble	0,7031	0,0444	0,1773	0,5485*	0,6926

* *Balanced Accuracy* estimada com base no *Recall* e proporção de classes

4.2.3 Considerações Finais

A abordagem Ensemble mostrou-se vantajosa ao equilibrar as métricas de avaliação, particularmente no que tange ao *F2-Score*, indicador especialmente relevante para

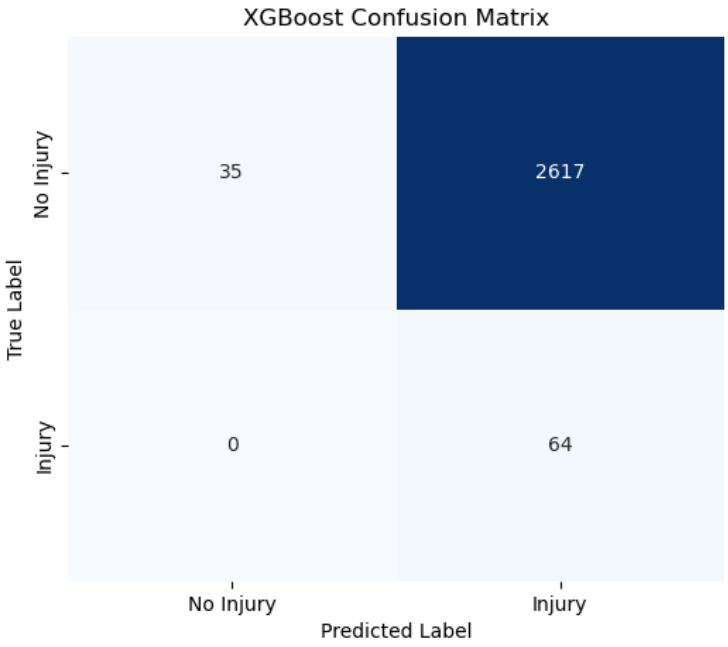

Figura 6 – Matriz de confusão do *XGBoost*.

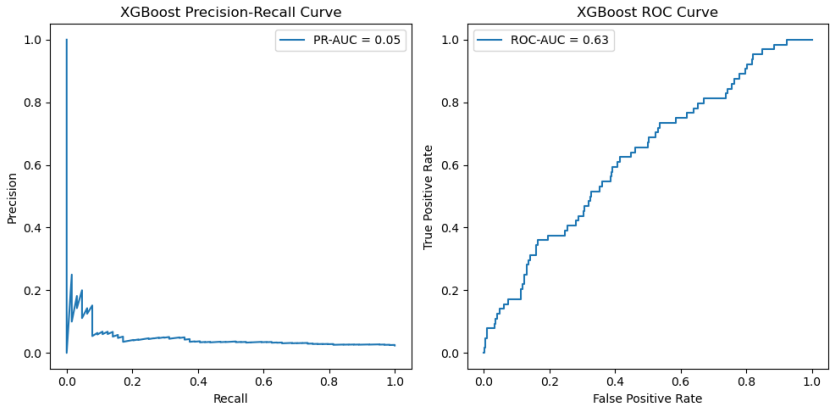

Figura 7 – Curvas PR e ROC do *XGBoost* (AUC-PR=0,0610, AUC-ROC=0,6207).

problemas com classes desbalanceadas. A combinação estratégica de modelos heterogêneos permitiu capturar padrões complementares, resultando em um classificador mais robusto e com melhor desempenho global quando comparado aos modelos individuais.

O *Ensemble* conseguiu manter um *Recall* competitivo enquanto significativamente melhorou a precisão, demonstrando a eficácia da abordagem de combinação de modelos para o problema em questão.

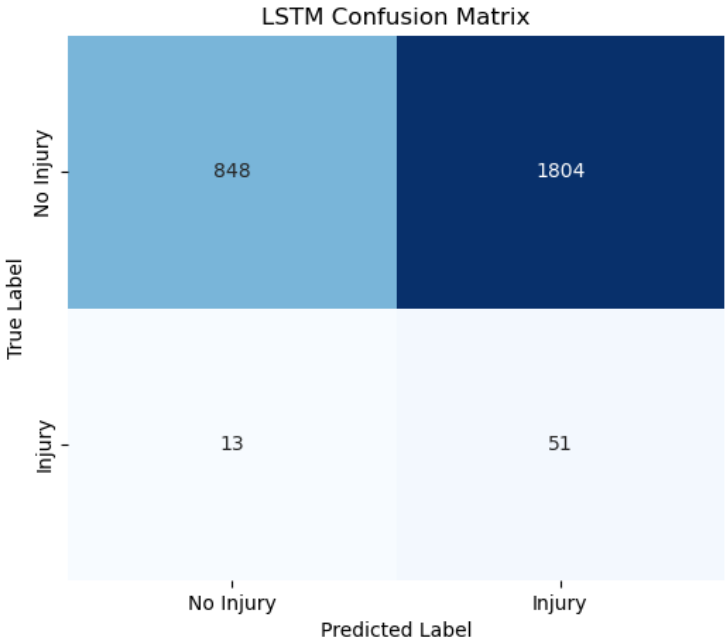

Figura 8 – Matriz de confusão do LSTM.

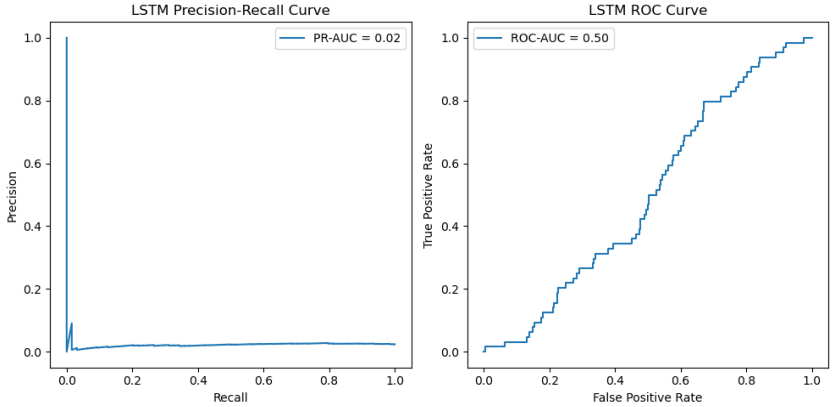

Figura 9 – Curvas PR e ROC do LSTM (AUC-PR=0,0340, AUC-ROC=0,5352).

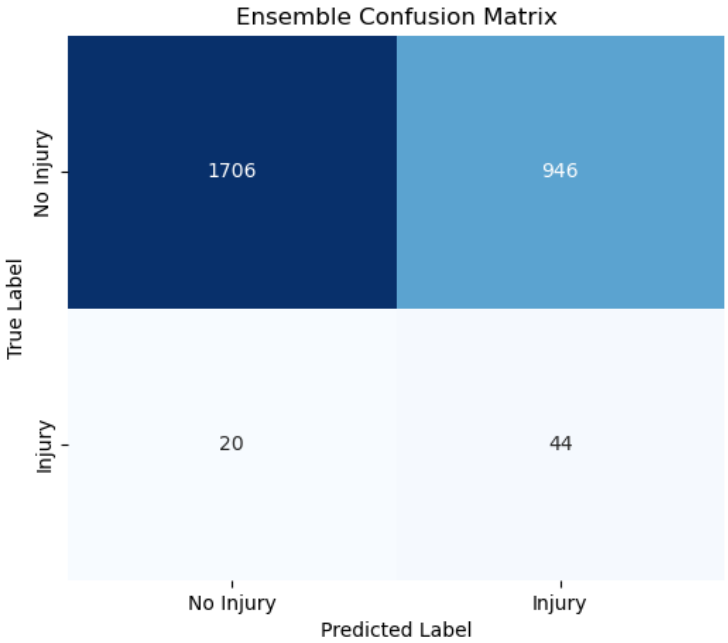

Figura 10 – Matriz de confusão do *Ensemble*.

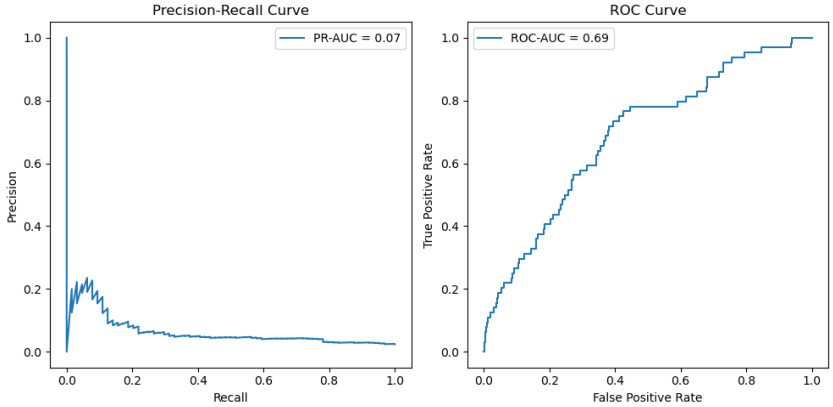

Figura 11 – Curvas PR e ROC do Ensemble (AUC-PR=0,0666, AUC-ROC=0,6926).

5 CONCLUSÕES

Este trabalho teve como objetivo desenvolver e avaliar um sistema preditivo de lesões em corredores de alto rendimento, utilizando técnicas de aprendizado de máquina e aprendizado profundo. A seguir, são apresentadas as principais conclusões obtidas a partir dos resultados.

5.1 Desempenho dos Modelos

Foram comparados quatro modelos individuais (Regressão Logística, *Random Forest*, *XGBoost* e LSTM) com um modelo *Ensemble*. Os resultados demonstraram que:

- O modelo LSTM apresentou o melhor desempenho entre os modelos individuais, com *F2-Score* de 0,1208 e Balanced Accuracy de 0,5536, destacando-se pelo alto *Recall* (0,7552), embora com baixa precisão (0,0279).
- O modelo *Ensemble* superou todos os demais, com *F2-Score* de 0,1773 — um aumento de 46,8% em relação ao LSTM — combinando *Recall* elevado (0,7031) com a maior precisão observada (0,0444).
- A abordagem *Ensemble* mostrou-se mais robusta, ao combinar as forças dos modelos base e melhorar o desempenho geral do sistema.

5.2 Importância das *Features*

A análise das variáveis mais relevantes para a predição de lesões revelou os seguintes destaques:

- Padrões temporais mostraram-se mais informativos do que valores pontuais, com médias móveis de 3 e 7 dias (sufixos *ma3* e *ma7*) aparecendo entre as *feature*s mais importantes.
- Volume e intensidade do treinamento, especialmente as variáveis *total_km_ma3* e *km_Z3-4_ma7*, foram identificadas como preditores relevantes.
- Percepções subjetivas dos atletas, como esforço percebido e recuperação, destacaram-se como variáveis altamente preditivas, reforçando o papel da autoavaliação no monitoramento esportivo.

5.3 Desafios do *Dataset* Desbalanceado

Um dos principais desafios enfrentados foi o forte desbalanceamento do conjunto de dados:

- Apenas cerca de 1,34% dos registros correspondiam a casos positivos (lesões), o que dificultou o treinamento dos modelos e afetou negativamente métricas como precisão.
- Apesar da aplicação de técnicas de balanceamento, a baixa frequência de lesões limitou a capacidade de generalização do sistema.

5.4 Escolha de Métricas e Abordagem

Dado o contexto médico-esportivo, as decisões metodológicas foram guiadas por critérios práticos:

- O *Recall* foi priorizado em relação à precisão, por se considerar mais crítico não detectar uma lesão iminente do que gerar um alerta falso.
- O *F2-Score* foi adotado como métrica principal, por valorizar mais o *Recall*, alinhando-se aos objetivos clínicos do sistema.
- Todas as escolhas de modelagem buscaram maximizar o impacto prático na prevenção de lesões e na segurança dos atletas.

5.5 Contribuições e Implicações Práticas

Este estudo gerou diversas contribuições relevantes:

- Validação da abordagem *Ensemble* como alternativa eficaz na predição de lesões em corredores.
- Identificação de variáveis críticas que podem subsidiar a tomada de decisões em estratégias de prevenção e periodização de treinos.
- Desenvolvimento de uma metodologia robusta para a construção de *Features* temporais aplicáveis a problemas semelhantes.
- Proposição de um *Framework* replicável com potencial de ser utilizado por treinadores, atletas e profissionais da saúde esportiva.

5.6 Limitações e Trabalhos Futuros

O presente estudo apresenta algumas limitações que podem orientar pesquisas futuras. A principal delas é o desbalanceamento extremo do conjunto de dados, composto majoritariamente por atletas de alto rendimento holandeses. A ampliação da base de dados com exemplos adicionais de lesões e a inclusão de atletas amadores — nos níveis iniciante e intermediário — podem aumentar a representatividade e a capacidade de generalização do modelo.

A integração de dados fisiológicos e biomecânicos representa outro caminho promissor, enriquecendo o modelo com informações complementares sobre carga, fadiga e recuperação. Além disso, a coleta de dados em plataformas populares, como o Strava, por meio de APIs, permitiria incorporar variáveis temporais, métricas de esforço percebido e padrões de recuperação, bem como dados ambientais e biométricos provenientes de outros aplicativos. Essa abordagem viabilizaria a validação do modelo em cenários mais diversos, aproximando-o de aplicações reais no contexto da saúde esportiva.

Paralelamente, o desenvolvimento de interfaces práticas para treinadores e atletas facilitaria o uso cotidiano do sistema, ampliando seu impacto no monitoramento preventivo. Também é recomendável a exploração de novas técnicas de aprendizado e balanceamento de classes para aprimorar o desempenho frente à baixa incidência de lesões.

Os resultados obtidos demonstram que técnicas de aprendizado de máquina, especialmente modelos Ensemble, são eficazes na predição de lesões em corredores. As variáveis mais relevantes identificadas fornecem insights valiosos para o planejamento do treinamento, destacando a importância do controle de carga, da recuperação adequada e da escuta das percepções dos próprios atletas.

A priorização do Recall como métrica principal reflete a preocupação com a segurança, reconhecendo que deixar de prever uma lesão é mais grave do que gerar um falso alerta.

Em síntese, este trabalho reforça o potencial da inteligência artificial aplicada à medicina esportiva, oferecendo uma base concreta para o desenvolvimento de ferramentas preditivas que promovam a longevidade e a segurança atlética.

REFERÊNCIAS

ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. **Statistics surveys**, v. 4, p. 40–79, 2010.

BAHR, R. Why we should focus on the burden of injuries and illnesses, not just their incidence. **British journal of sports medicine**, BMJ Publishing Group Ltd, v. 46, n. 1, p. 10–12, 2012.

BLANCH, P.; GABBETT, T. J. Has the athlete trained enough to return to play safely? the acute: chronic workload ratio permits clinicians to quantify a player's risk of subsequent injury. **British journal of sports medicine**, BMJ Publishing Group Ltd, v. 50, n. 8, p. 471–475, 2016.

BOURDON, P. C. *et al.* Monitoring athlete training loads: consensus statement. **International journal of sports physiology and performance**, Human Kinetics, v. 12, n. S2, p. S2–161–S2–170, 2017.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. Disponível em: https://link.springer.com/article/10.1023/A:1010933404324.

BRODERSEN, K. H. *et al.* The balanced accuracy and its posterior distribution. **2010 20th international conference on pattern recognition**, IEEE, p. 3121–3124, 2010.

BUIST, I. *et al.* Incidence and risk factors of running-related injuries during preparation for a 4-mile recreational running event. **British journal of sports medicine**, BMJ Publishing Group Ltd, v. 44, n. 8, p. 598–604, 2010.

CAREY, D. L. *et al.* Predictive modelling of training loads and injury in australian football. **Journal of Science and Medicine in Sport**, v. 20, n. 2, p. 165–170, 2017.

CHAWLA, N. V. *et al.* Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. *In*: ACM. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [*S.l.: s.n.*], 2016. p. 785–794.

CHINCHOR, N. Muc-4 evaluation metrics. *In*: **Proceedings of the 4th conference on Message understanding**. [*S.l.: s.n.*], 1992. p. 22–29.

CLAUDINO, J. G. *et al.* The role of data analytics in sports injury prevention. **Frontiers in Physiology**, v. 10, p. 1354, 2019.

COLBY, M. J. *et al.* Accelerometer and gps-derived running loads and injury risk in elite australian footballers. **The Journal of Strength & Conditioning Research**, LWW, v. 31, n. 12, p. 3604–3614, 2017.

DAVIS, J.; GOADRICH, M. The relationship between precision-recall and roc curves. **Proceedings of the 23rd international conference on Machine learning**, p. 233–240, 2006.

FLECK, S. J.; KRAEMER, W. J. **Designing Resistance Training Programs**. Champaign, IL: Human Kinetics, 2014.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of computer and system sciences**, Elsevier, v. 55, n. 1, p. 119–139, 1997.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001.

GABBETT, T. J. The training-injury prevention paradox: Should athletes be training smarter and harder? **British Journal of Sports Medicine**, v. 50, n. 5, p. 273–280, 2016.

GENT, R. N. van *et al.* Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review. **British journal of sports medicine**, BMJ Publishing Group Ltd, v. 41, n. 8, p. 469–480, 2007.

GRAVES, A. Supervised sequence labelling with recurrent neural networks. **Studies in Computational Intelligence**, Springer, v. 385, p. 5–13, 2012.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of machine learning research**, v. 3, n. Mar, p. 1157–1182, 2003.

HAN, H.; WANG, W.-Y.; MAO, B.-H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. **International conference on intelligent computing**, Springer, p. 878–887, 2005.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. New York, NY: Springer, 2009.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. [*S.l.: s.n.*]: Springer Science & Business Media, 2009.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on knowledge and data engineering**, IEEE, v. 21, n. 9, p. 1263–1284, 2009.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, 1997.

HUBER, P. J.; RONCHETTI, E. M. **Robust statistics**. [*S.l.: s.n.*]: John Wiley & Sons, 2011.

HULIN, B. T. *et al.* The acute: chronic workload ratio predicts injury: high chronic workload may decrease injury risk in elite rugby league players. **British journal of sports medicine**, BMJ Publishing Group Ltd, v. 50, n. 4, p. 231–236, 2014.

IMPELLIZZERI, F. M. *et al.* Use of rpe-based training load in soccer. **Medicine and science in sports and exercise**, v. 36, n. 6, p. 1042–1047, 2004.

KELLMANN, M. *et al.* Recovery and performance in sport: consensus statement. **International journal of sports physiology and performance**, Human Kinetics, v. 13, n. 2, p. 240–245, 2018.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical analysis with missing data**. [*S.l.: s.n.*]: John Wiley & Sons, 2019.

LOVDAL, S.; HARTIGH, R. J. R. D.; AZZOPARDI, G. Injury prediction in competitive runners with machine learning. **International Journal of Sports Physiology and Performance**, in press, 2020.

LUGER, G. F. **Inteligência Artificial: Estruturas e Estratégias para Solução Complexa de Problemas**. São Paulo: Pearson, 2017.

PROVOST, F.; FAWCETT, T.; KOHAVI, R. The case against accuracy estimation for comparing induction algorithms. **Proceedings of the Fifteenth International Conference on Machine Learning**, v. 98, p. 445–453, 2001.

ROKACH, L. Ensemble-based classifiers. **Artificial Intelligence Review**, Springer, v. 33, n. 1, p. 1–39, 2010.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 8, n. 4, p. e1249, 2018.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. **PloS one**, Public Library of Science, v. 10, n. 3, p. e0118432, 2015.

SARAGIOTTO, B. T. *et al.* What are the main risk factors for running-related injuries? **Sports medicine**, Springer, v. 44, n. 8, p. 1153–1163, 2014.

SEILER, S. What is best practice for training intensity and duration distribution in endurance athletes? **International Journal of Sports Physiology and Performance**, v. 5, n. 3, p. 276–291, 2010.

SHERSTINSKY, A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. **Physica D: Nonlinear Phenomena**, Elsevier, v. 404, p. 132306, 2020.

SOLIGARD, T. *et al.* 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**, BMJ Publishing Group Ltd, v. 50, n. 17, p. 1030–1041, 2016.

TOLEDO, M. **CNN Brasil**. 2024. Disponível em: https://www.cnnbrasil.com.br/saude/corrida-foi-o-esporte-mais-praticado-no-mundo-em-2024-diz-relatorio-veja-dados/.

WOLPERT, D. H. Stacked generalization. **Neural Networks**, v. 5, n. 2, p. 241–259, 1992.

ZINNER, C.; SPERLICH, B. Exercise and the risk of injuries in athletes. **Journal of Sports Science and Medicine**, v. 14, p. 1–9, 2015. Disponível em: https://www.jssm.org/.