CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

ANÁLISE DE DESEMPENHO E CARACTERÍSTICAS DE CORREDORES PARA PREDIÇÃO DE RESULTADOS E GERAÇÃO DE FEEDBACKS PERSONALIZADOS

Nathália Carvalho Tito

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Glauco Fiorott Amorim
Eduarado Bezerra da Silva

Rio de Janeiro,
27 de Abril de 2026

Análise de Desempenho e Características de Corredores para Predição de Resultados e Geração de Feedbacks Personalizados

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Nathália Carvalho Tito

Aprovada por:

Presidente, Prof. Glauco Fiorott Amorim, D.Sc. (orientador)

Prof. Eduardo Bezerra da Silva, D.Sc. (coorientador)

Prof. Diego Nunes Brandão, D.Sc.

Prof. Cláudio Miceli de Farias, D.Sc.

Rio de Janeiro,

27 de Abril de 2026

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

T621 Tito, Nathália Carvalho
Análise de desempenho e características de corredores para predição de resultados e geração de feedbacks personalizados / Nathália Carvalho Tito. — 2026.
81f. : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2026.
Bibliografia : f. 76-81
Orientador: Glauco Fiorott Amorim
Coorientador: Eduardo Bezerra da Silva

1. Corridas (Atletismo) - Análise. 2. Aprendizado do computador. 3. Análise por agrupamento. 4. Corredores (Esportes). I. Amorim, Glauco Fiorott. (Orient.). II. Silva, Eduardo Bezerra da. (Coorient.). III. Título.

CDD 796.425

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

DEDICATÓRIA

Dedico este trabalho aos meus pais, cujo amor, apoio constante e ensinamentos tornaram possível cada passo desta jornada. Aos meus amigos, que estiveram presentes com incentivo, parceria e acolhimento nos momentos desafiadores e nos de celebração, deixo minha gratidão mais sincera. Este trabalho também é de vocês.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, pela força, proteção e sabedoria concedidas ao longo de todo este percurso. Expresso minha profunda gratidão aos meus pais, pelo amor incondicional, pelo incentivo constante e pelo exemplo de dedicação que sempre me inspirou. Aos meus professores, agradeço pela orientação, pelos ensinamentos fundamentais e pelo compromisso com minha formação acadêmica. Estendo minha gratidão aos meus amigos, que ofereceram apoio, companhia e palavras de encorajamento nos momentos de maior desafio. Registro também meu reconhecimento aos meus líderes profissionais, cuja confiança, orientação e experiências compartilhadas contribuíram significativamente para meu crescimento técnico e pessoal. Por fim, agradeço ao órgão de fomento, responsável por fomentar esta pesquisa.

RESUMO

Análise de Desempenho e Características de Corredores para Predição de Resultados e Geração de Feedbacks Personalizados

Nathália Carvalho Tito

Orientadores:
Glauco Fiorott Amorim
Eduarado Bezerra da Silva

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

O aumento do número de praticantes de corrida de rua tem intensificado a demanda por soluções capazes de oferecer suporte individualizado ao treinamento, especialmente entre corredores amadores que carecem de acompanhamento profissional contínuo. Nesse contexto, este estudo propõe um modelo integrado de análise de desempenho baseado em técnicas de aprendizado de máquina, com o objetivo de compreender padrões de treino, prever o desempenho de corredores e gerar recomendações personalizadas a partir de variáveis controláveis. Foram utilizados dados provenientes de um questionário e de registros de atividades exportados de um aplicativo de corrida, envolvendo 26 atletas com diferentes níveis de experiência, permitindo uma visão multidimensional dos hábitos de treino e da experiência esportiva.

A metodologia foi estruturada em três fases principais. Na primeira, realizou-se uma análise de clusterização utilizando algoritmos clássicos (*K-means*, *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) e *Agglomerative Clustering*) aplicados a componentes principais responsáveis por 80% da variabilidade dos dados. O modelo final selecionado identificou três perfis distintos: "Experientes Jovens", "Pouco Experientes" e "Experientes Veteranos", diferenciados por idade, maturidade esportiva e padrões de treino.

Na segunda fase, desenvolveu-se um modelo preditivo baseado em *gradient boosting*, utilizando o algoritmo *Extreme Gradient Boosting* (XGBoost), tanto em uma configuração geral quanto em versões específicas para cada *cluster*. Modelos de Regressão Linear também foram testados como abordagem de referência; entretanto, o XGBoost apresentou desempenho superior em todos os cenários avaliados, evidenciando maior capacidade de capturar relações não lineares e interações complexas entre as variáveis. Os resultados indicaram ainda que cada grupo respondeu de maneira particular às variáveis de treino, reforçando que a segmentação aprimora substancialmente a capacidade preditiva e a adequação das recomendações.

A interpretação dos modelos foi conduzida por meio da análise dos *Shapley Additive Explanation* (SHAP) *values*, que permitiu identificar as variáveis de maior influência nas previsões. De forma geral, destacaram-se fatores diretamente relacionados ao desempenho e à estrutura do treino, como velocidades mínima e máxima, distância percorrida, variabilidade do ritmo e histórico recente de velocidade, além de variáveis contextuais como temperatura. A análise segmentada revelou padrões distintos entre os grupos, indicando que diferentes perfis de corredores apresentam sensibilidades

específicas a aspectos como intensidade, regularidade e consistência do treinamento, o que reforça a importância da personalização das recomendações.

Na terceira fase, aplicou-se um algoritmo de otimização para identificar combinações de variáveis controláveis capazes de maximizar o desempenho previsto em cada perfil. A avaliação dos resultados foi fundamentada na comparação com cenários *baseline*, nos quais não havia ajuste das variáveis de decisão, permitindo quantificar de forma objetiva os ganhos obtidos com a otimização. As recomendações derivadas mostraram coerência interna e alinhamento com características dos *clusters*: maior ênfase em intensidade e variedade para corredores jovens experientes; foco em regularidade, consistência e fortalecimento muscular para pouco experientes; e estratégias de manutenção, equilíbrio e controle de carga para veteranos. Os achados demonstram que a integração entre clusterização, modelagem preditiva e otimização forma uma abordagem consistente e promissora para o desenvolvimento de sistemas inteligentes de recomendação esportiva baseados em dados.

Apesar das limitações relacionadas ao tamanho da amostra e à ausência de indicadores fisiológicos de maior granularidade, o estudo apresenta evidências iniciais de que modelos computacionais podem apoiar a personalização do treinamento de forma eficiente, acessível e escalável. Futuras pesquisas podem ampliar a base de dados, incorporar novas dimensões informacionais, validar o modelo em populações maiores e explorar sua aplicação em plataformas digitais. Conclui-se que a combinação de técnicas de ciência de dados e métodos de otimização contribui significativamente para a compreensão do desempenho em corrida e para o desenvolvimento de recomendações individualizadas que favorecem evolução, aderência e segurança na prática esportiva.

Palavras-chave: Análise esportiva; Corrida; Análise de desempenho; Aprendizado de máquina; Clusterização; Modelagem preditiva; Otimização; Recomendações personalizadas; Perfis de corredores.

Rio de Janeiro,

27 de Abril de 2026

ABSTRACT

Análise de Desempenho e Características de Corredores para Predição de Resultados e Geração de Feedbacks Personalizados

Nathália Carvalho Tito

Advisors:
Glauco Fiorott Amorim
Eduarado Bezerra da Silva

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

The increasing number of recreational runners has intensified the demand for solutions capable of providing individualized training support, especially among amateur athletes who often lack continuous professional guidance. In this context, this study proposes an integrated performance analysis model based on machine learning techniques, aiming to understand training patterns, predict runners' performance, and generate personalized recommendations based on controllable variables. Data were obtained from a questionnaire and activity records exported from a running application, involving 26 athletes with different levels of experience, allowing a multidimensional view of training habits and sports experience.

The methodology was structured in three main phases. In the first phase, a clustering analysis was performed using classical algorithms (K-means, DBSCAN, and Agglomerative Clustering) applied to principal components explaining 80% of the data variability. The selected model identified three distinct profiles: "Young Experienced," "Less Experienced," and "Veteran Experienced," differentiated by age, sports maturity, and training patterns.

In the second phase, a predictive model based on gradient boosting was developed using the XGBoost algorithm, both in a general configuration and in versions specific to each cluster. Linear regression models were also tested as a reference approach; however, XGBoost achieved superior performance across all evaluated scenarios, demonstrating a greater ability to capture nonlinear relationships and complex interactions among variables. The results also indicated that each group responded differently to training variables, reinforcing that segmentation substantially improves predictive performance and the adequacy of personalized recommendations.

Model interpretability was investigated using SHAP values, which enabled the identification of the most influential variables in the predictions. In general, variables directly related to training structure and performance stood out, such as minimum and maximum speed, distance covered, pace variability, and recent performance history, as well as contextual factors such as temperature. The segmented analysis revealed distinct patterns across groups, indicating that different runner profiles present specific sensitivities to aspects such as intensity, regularity, and training consistency, further supporting the importance of personalized recommendations.

In the third phase, an optimization algorithm was applied to identify combinations of controllable variables capable of maximizing the predicted performance for each profile. The evaluation of results was based on comparisons with baseline scenarios, in which decision variables were not

adjusted, allowing objective quantification of the gains achieved through optimization. The resulting recommendations showed internal coherence and alignment with cluster characteristics: greater emphasis on intensity and training variety for young experienced runners; focus on regularity, consistency, and strength training for less experienced runners; and strategies centered on maintenance, balance, and load control for veteran runners. These findings demonstrate that the integration of clustering, predictive modeling, and optimization provides a consistent and promising approach for developing data-driven intelligent sports recommendation systems.

Despite limitations related to sample size and the absence of more granular physiological indicators, the study provides initial evidence that computational models can effectively support personalized training in an efficient, accessible, and scalable manner. Future research may expand the dataset, incorporate additional informational dimensions, validate the model in larger populations, and explore its implementation in digital platforms. In conclusion, the combination of data science techniques and optimization methods significantly contributes to understanding running performance and to developing individualized recommendations that promote improvement, adherence, and safety in sports practice.

Key-words: Sports analysis; Running; Performance analysis; Machine learning; Clustering; Predictive modeling; Optimization; Personalized recommendations; Runner profiles.

Rio de Janeiro,

27 de Abril de 2026

Sumário

I Introdução **1**

I.1 Contextualização 1

I.2 Motivação 2

I.3 Objetivos e Contribuições da Pesquisa 5

II Referencial Teórico **8**

II.1 Conceitos de Séries Temporais 8

II.2 Algoritmos de Clusterização 9

II.2.1 *K-means* 9

II.2.2 *Aglomerative Hierarchical Clustering* 10

II.2.3 DBSCAN 10

II.3 Modelagem Preditiva 11

II.3.1 Regressão Linear 11

II.3.2 XGBoost 12

II.4 Algoritmos Otimizadores 13

III Trabalhos Relacionados **16**

III.1 Séries Temporais no Esporte 16

III.2 Análise de *Cluster* no Esporte 17

III.3 Previsão de Desempenho na Corrida 19

III.4 Otimização de Variáveis Para Maximizar um Objetivo 23

III.5 Recomendações Personalizadas 24

III.6 Posicionamento do Estudo 27

IV Metodologia **30**

IV.1 Coleta dos Dados 30

IV.2 Pré-processamento dos Dados 32

IV.2.1 Geração e Enriquecimento de Variáveis 32

IV.2.2 Limpeza dos Dados 38

IV.3 Modelagem: As 3 Fases 39

IV.3.1 Clusterização de Corredores 39
IV.3.2 Previsão de Desempenho 42
IV.3.3 Recomendação para Maximização do Desempenho 47

V Resultados 50
V.1 Resultados da Pesquisa 50
V.1.1 Base Questionário 50
V.1.2 Base Atividades 53
V.2 Resultados da Modelagem 54
V.2.1 Fase 1 - Encontrando Perfil dos Corredores 55
V.2.2 Fase 2 - Prevendo Desempenho dos Corredores 57
V.2.3 Fase 3 - Recomendação Personalizada por Perfil 64

VI Considerações Finais 69
VI.1 Principais Limitações 69
VI.2 Evoluções e Trabalhos Futuros 71
VI.3 Conclusões 72

Referências 76

Lista de Figuras

I.1 Série temporal de velocidade durante um treino de corrida de um corredor 4

IV.1 Fluxo resumido para obtenção dos dados 32
IV.2 Fluxo de treinamento dos modelos 40
IV.3 Distribuição dos indicadores de treinamento 44

V.1 Distribuição dos meses de prática dos lesionados e não lesionados 52
V.2 Distribuição da velocidade média por períodos de lesão 53
V.3 Distribuição da velocidade média por frequência de prática de outros exercícios 54
V.4 Variabilidade capturada por componente 55
V.5 Dendograma - Agglomerative Clustering 56
V.6 Distribuição do desvio padrão da velocidade 58
V.7 SHAP *values* - Contribuição das variáveis (Geral) 59
V.8 SHAP *values* - Contribuição das variáveis (*Cluster* 0) 61
V.9 SHAP *values* - Contribuição das variáveis (*Cluster* 1) 62
V.10 SHAP *values* - Contribuição das variáveis (*Cluster* 2) 63

VI.1 Fluxo de análise de desempenho e recomendação 71

Lista de Tabelas

III.1 Estudos de previsão de desempenho na corrida 21

III.2 Contribuições do presente estudo em relação à literatura 28

IV.1 Variáveis originais extraídas do questionário e do histórico de atividades 33

IV.2 Variáveis derivadas criadas para análise 37

IV.3 Resultados dos testes dos indicadores de treino 44

V.1 Resultado da otimização da função objetivo (velocidade média prevista) - Geral 65

V.2 Resultado da otimização da função objetivo (velocidade média prevista) - *Cluster* 0 67

V.3 Resultado da otimização da função objetivo (velocidade média prevista) - *Cluster* 1 67

V.4 Resultado da otimização da função objetivo (velocidade média prevista) - *Cluster* 2 67

V.5 Comparação entre previsões otimizadas e *baseline* 68

Lista de Acrônimos

ANN *Artificial Neural Network*

ARMA *Autoregressive Moving Average*

BFGS *Broyden–Fletcher–Goldfarb–Shanno*

BLUE *Best Linear Unbiased Estimator*

CAAE Certificado de Apresentação para Apreciação Ética

CBR Raciocínio Baseado em Casos

COBYLA *Constrained Optimization by Linear Approximations*

DBSCAN *Density-Based Spatial Clustering of Applications with Noise*

EMGEC Computação Evolutiva Multi-Gradiente Extrema

GRU *Gated Recurrent Unit*

IFRJ Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro

IGWO Otimização do Lobo Cinzento Aprimorada

IMC Índice de massa corporal

KNN *K-Nearest Neighbors*

L-BFGS-B *Limited-memory Broyden–Fletcher–Goldfarb–Shanno - Bounded*

La$^-$bAT Concentração de lactato no sangue no limiar anaeróbico

La$^-$bRCP Concentração de lactato no sangue no ponto de compensação respiratória

LSTM *Long Short-Term Memory*

MIO Otimização Inteira Mista

MOMS *Motivation of Marathoners Scale*

OAT *one-at-a-time*

OLS *Ordinary Least Squares*

PCA *Principal Component Analysis*

R^2 Coeficiente de Determinação

RCP Ponto de Compensação Respiratória

RMSE *Root Mean Squared Error*

RNN *Recurrent Neural Network*

SBBD Simpósio Brasileiro de Banco de Dados

SHAP *Shapley Additive Explanation*

SLSQP *Sequential Least Squares Programming*

TCLE Termo de Consentimento Livre e Esclarecido

VEmax Ventilação pulmonar máxima

VERCP Ventilação pulmonar no ponto de compensação respiratória

VO_2AT Consumo de oxigênio no limiar anaeróbico

VO_2RCP Consumo de oxigênio no ponto de compensação respiratória

VO_2max Consumo máximo de oxigênio

XGBoost *Extreme Gradient Boosting*

WCSS *Within-Cluster Sum of Squares*

Capítulo I Introdução

I.1 Contextualização

A corrida tem apresentado um crescimento bastante expressivo no Brasil. Entre 2007 e 2015, o número de corridas de rua em cidades como Rio de Janeiro e São Paulo aumentou em mais de 90%, com tendência linear positiva e significativa de participação em ambos os sexos [Thuany et al., 2021a]. Na tradicional Corrida de São Silvestre, por exemplo, verificou-se que a participação de corredores cresceu ao longo dos anos para homens e mulheres, embora acompanhada por uma queda geral no desempenho. Além disso, foram observadas diferenças significativas entre os sexos, com os homens apresentando melhores tempos, e um aumento progressivo nessa disparidade de desempenho [Thuany et al., 2023]. Em estudo semelhante sobre a Maratona de Berlim, identificou-se também crescimento na participação de homens e mulheres, mas com maior aumento entre as mulheres, o que reduziu a desigualdade na participação. Contudo, de forma consistente, os desempenhos médios diminuíram em todas as faixas etárias [Reusser et al., 2021].

Com a crescente popularização da corrida, observa-se também maior mobilidade geográfica dos atletas em busca de eventos esportivos. O estudo de Thuany et al. [2021b], ao analisar o fenômeno denominado *OUTrun* (probabilidade de sair do estado de residência para participar de provas em outro estado ou país), identificou que as regiões Sul e Sudeste do Brasil concentram o maior número de participantes, enquanto corredores do Norte e Nordeste são os que mais se deslocam para competir. Os autores ainda apontam que fatores como distância preferida, nível de renda e idade estão associados a esse comportamento.

Além do aspecto geográfico, múltiplas motivações levam indivíduos a ingressar na prática da corrida. Zach et al. [2017] propuseram um modelo detalhado das motivações para corredores de maratona por meio da *Motivation of Marathoners Scale* (MOMS). De acordo com os autores, uma escala revisada identifica 11 fatores motivacionais, organizados em dimensões como gestão do estresse (incluindo gestão emocional e do cotidiano), significado de vida, autoestima, saúde geral (com subdivisões entre reduzir doenças e manter a forma), preocupação com o peso, reconhecimento, afiliação, competição e realização de metas pessoais. Concluem ainda que essa estrutura multifacetada se alinha com a Teoria da Autodeterminação, na medida em que as motivações mapeadas refletem a busca pela satisfação das necessidades de autonomia, competência e relacionamento.

O modelo multidimensional de motivação proposto por Zach et al. [2017] corrobora as conclusões de Parra-Camacho et al. [2019], ao demonstrar empiricamente que a participação e a adesão à corrida são influenciadas por um complexo espectro de fatores, que incluem desde razões de natureza psicológica e social até aspirações competitivas e de gestão do tempo. Entre os fatores que mais impulsionam a participação e a permanência na corrida, segundo estudo de Koronios et al. [2017], estão aqueles associados à motivação intrínseca, como o prazer, a satisfação pessoal e a superação de desafios, que refletem o desejo de aprimoramento individual, o sentimento de competência e a autonomia proporcionada pela prática. Por outro lado, a motivação extrínseca também desempenha papel relevante, englobando recompensas sociais, reconhecimento, e a busca por pertencimento em comunidades de corredores. De acordo com a teoria da autodeterminação, a continuidade na prática tende a ocorrer quando essas necessidades psicológicas fundamentais (autonomia, competência e relacionamento) são atendidas, favorecendo o engajamento duradouro e a intenção de continuar correndo [Koronios et al., 2017].

Como efeito desse crescimento, a corrida tem se mostrado uma aliada na promoção da qualidade de vida, impactando positivamente diversos indicadores biomédicos de saúde [Hespanhol Junior et al., 2015]. Esse benefício foi especialmente observado em praticantes de longa distância, que não apenas relataram melhorias significativas na saúde mental (93% dos entrevistados), mas também adotaram hábitos alimentares conscientes, refletindo uma busca por um estilo de vida mais saudável e sustentável. Além disso, a utilização de tecnologias vestíveis para monitoramento de parâmetros como frequência cardíaca, também indica uma tendência crescente de autocuidado e gestão proativa da saúde entre corredores [Lundy and Reilly, 2024].

I.2 Motivação

O desempenho em corridas e, de modo mais amplo, o alcance de metas esportivas específicas é resultado de uma interação complexa entre múltiplos fatores. Pesquisas na área indicam que variáveis relacionadas às características individuais dos corredores, aos aspectos do treinamento e a condições externas exercem influência significativa sobre os resultados obtidos. Nesse contexto, o sucesso esportivo pode não depender apenas do esforço individual, mas de um conjunto integrado de fatores fisiológicos, comportamentais e ambientais que, em conjunto, determinam o desempenho final.

Segundo V. et al. [2022], o resultado de um atleta na corrida depende de fatores como sexo, massa muscular, tempo de corrida, frequência cardíaca, pulso, velocidade, precisão e resistência. Características demográficas, por exemplo, podem ser utilizadas na predição do tempo de prova [Lerebourg et al., 2023], assim como a intensidade do treino, treino de distâncias mais longas, melhor qualidade do sono e concentração de lactato sanguíneo, também se mostram relevantes

nessa descoberta [Shu et al., 2024; Tanda, 2011; Knechtle et al., 2011; Pérez et al., 2012].

Quando entra-se no âmbito de dados passados, em um contexto temporal, treinos antigos podem ser modelados em séries temporais, onde modelos de aprendizado profundo podem atuar em previsões de desempenho de um corredor [El-Kassabi et al., 2020]. Fatores antropométricos, fisiológicos e biomecânicos contribuem adicionalmente para explicar o desempenho [Gómez-Molina et al., 2017]. Ainda considerando aspectos fisiológicos, a ventilação pulmonar (maior suprimento de oxigênio (O_2)) também se mostra associada ao melhor desempenho na corrida [Wiecha et al., 2022].

Adicionalmente, condições ambientais, como a temperatura, apresentam forte influência sobre a velocidade de corrida, conforme demonstrado por El Helou et al. [2012], que desenvolveram um modelo para relacionar desempenho à temperatura do ar, identificando as condições ideais para atingir metas. Compreender esses fatores e ter uma estimativa realista de onde o atleta pode chegar não apenas fornece subsídios para planejamento do treino, mas também atua como fator motivacional, incentivando o corredor a perseguir objetivos de forma mais estratégica e consistente. Corroborando essa perspectiva, Ely et al. [2008] examinaram o ritmo geral de mulheres maratonistas e avaliou a influência da temperatura do ar no ritmo de atletas com diferentes habilidades ao longo da prova. O estudo concluiu que o efeito da temperatura ambiente no ritmo depende da habilidade da atleta, sendo as mais lentas mais negativamente impactadas.

O processo de predição de desempenho pode ser desafiador, uma vez que os atletas apresentam diferentes perfis e, consequentemente, distintas estratégias de ritmo, que ainda podem variar conforme o tipo de prova [Casado et al., 2021]. Nesse contexto, segmentar e compreender as diversas comunidades de corredores torna-se fundamental para análises de desempenho baseadas em séries de velocidade. Por exemplo, Besomi et al. [2018] identificaram perfis de corredores urbanos por meio de análise de *cluster*, considerando dados sociodemográficos, de saúde, de motivação, características de treinamento e comportamentos relacionados à corrida, revelando perfis distintos com diferentes preferências. Além disso, Nikolaidis et al. [2017] investigaram o ritmo e a idade do pico de desempenho em maratonas, constatando que as mulheres alcançam seu melhor tempo aproximadamente cinco anos antes dos homens, ilustrando como o desempenho pode variar de acordo com múltiplos fatores.

A Figura I.1 [Tito et al., 2024] ilustra um exemplo de uma série temporal da velocidade, medida em km/h e a cada segundo, em um treino de corrida de um corredor, sugerindo como a velocidade pode oscilar dentro de um treino e como pode ser desafiador realizar análises e eventuais previsões. Nota-se que em alguns momentos do treino de corrida, o atleta diminui bruscamente sua velocidade, retomando em seguida para o patamar inicial. Esse evento pode ter ocorrido não só por motivos que envolvem experiência (hipótese principal do estudo) mas também por características físicas, biológicas, de treinamento, ou até mesmo por algum evento externo. Um exemplo de como variáveis

exógenas podem ser usadas para apoiar esse tipo de análise está no estudo de Gosztyla et al. [2006], que analisa o ritmo ideal para prova de 5 km e correlata diversas informações antropométricas e de oxigênio com a mudança de ritmo.

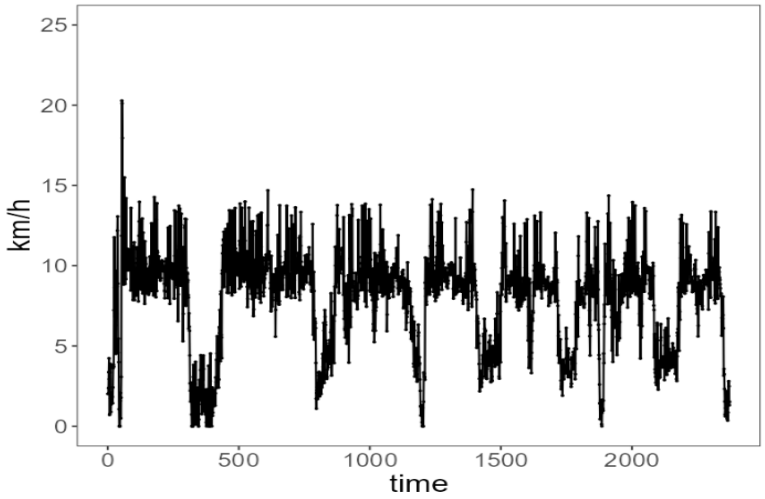


Figura I.1: Série temporal de velocidade durante um treino de corrida de um corredor

Apesar dos inúmeros benefícios associados à corrida, diversos fatores também podem atuar como barreiras à continuidade da prática. Koronios et al. [2017] demonstrou em seu estudo que entre esses principais fatores estão a limitação de tempo, falta de motivação, lesões, ausência de companhia para treinar e dificuldades em manter uma rotina regular de exercícios. Aspectos psicológicos, como o medo de não atingir expectativas de desempenho ou a percepção de baixo progresso, também podem contribuir para o desengajamento. Além disso, condições ambientais, financeiras e sociais podem restringir o acesso a espaços adequados para o treino ou à participação em eventos esportivos, reduzindo o estímulo à prática [Koronios et al., 2017]. Essas limitações, quando não acompanhadas de estratégias eficazes de monitoramento e suporte ao corredor, tendem a comprometer a constância do treinamento e o alcance de metas de desempenho, evidenciando a importância de abordagens que favoreçam a manutenção da prática e a motivação ao longo do tempo.

Monitorar o desenvolvimento do desempenho do corredor permite tomar decisões mais direcionadas para otimizar a estratégia e, ao mesmo tempo, atua como um estímulo à continuidade da prática. Segundo Komitova et al. [2022], a utilização de séries temporais apresenta grande potencial para compreender fenômenos relacionados ao desempenho, oferecendo métodos relevantes para apoiar avaliações e favorecer a criação de novas estratégias nas ciências do esporte. Com a crescente disponibilidade de dados temporais e o avanço no número de aplicativos, sensores e dispositivos voltados ao monitoramento da saúde e da atividade física, surge um campo significativo para a aplicação de técnicas de aprendizado de máquina na análise esportiva [Smyth, 2019].

Nesse contexto, recomendações adaptativas e personalizadas ao perfil de cada atleta durante a

prova poderiam orientar o corredor a alcançar melhor desempenho e favorecer o cumprimento de seus objetivos [Berndsen et al., 2019], além de estimular a manutenção dos treinos no esporte. A prática de atividade física, aliada ao uso de tecnologias que incentivem escolhas mais saudáveis, por meio de recomendações persuasivas e adaptadas a diferentes rotinas e hábitos, pode contribuir para transformar comportamentos nocivos em práticas benéficas, apoiando-se em uma gestão inteligente da carga de treinamento e do tempo de recuperação [Smyth, 2019]. Em corridas de longa distância, como maratona por exemplo, recomendações personalizadas com base em dados biológicos e ambientais são essenciais para prevenir sobrecarga, otimizar o desempenho e garantir a segurança do atleta, adaptando estratégias de forma dinâmica durante o percurso [V. et al., 2022].

É possível ampliar a discussão destacando como os sistemas de recomendação podem atuar de forma integrada ao longo de toda a jornada do corredor, desde o planejamento inicial até a recuperação pós-prova. Smyth et al. [2022] apresentam uma visão abrangente de como técnicas de aprendizado de máquina e sistemas de recomendação podem ser aplicadas em diferentes fases do treinamento, incluindo a estimativa de condicionamento físico personalizado, a recomendação de planos de treino adaptativos e a previsão de risco de lesões com base na carga de treinamento. Essas abordagens permitem que corredores recreacionais, muitas vezes sem acesso a treinadores profissionais, recebam orientações precisas e contextualizadas, melhorando não apenas seu desempenho, mas também sua segurança e adesão ao esporte.

Além disso, os autores destacam oportunidades secundárias, porém igualmente relevantes, de recomendação no contexto da maratona, como a sugestão de eventos complementares, parceiros de treino, rotas, equipamentos adequados e até mesmo conteúdo motivacional [Smyth et al., 2022]. Essas recomendações auxiliam na manutenção do engajamento do corredor ao longo de meses de preparação, um período que frequentemente envolve desafios físicos e psicológicos. A integração dessas diferentes dimensões (treino, recuperação, nutrição e motivação) em uma única plataforma inteligente pode potencializar os resultados do atleta, criando um ecossistema de suporte contínuo e altamente personalizado, que se adapta às suas necessidades em tempo real [Smyth et al., 2022].

I.3 Objetivos e Contribuições da Pesquisa

Com a capacidade de entender como se dá o desempenho de um atleta, decisões otimizadas podem ser tomadas para eventuais ajustes de estratégia, o que poderia acontecer através de recomendações e *feedbacks* interativos. Esse sistema beneficia principalmente atletas amadores que não tem acesso a acompanhamento profissional, já que teriam uma espécie de treino personalizado para o seu perfil, dado o seu objetivo [Berndsen et al., 2017].

Nesse contexto, o objetivo principal deste trabalho consiste em desenvolver uma modelagem preditiva personalizada capaz de estimar o desempenho de corredores, considerando características

individuais, hábitos e histórico de treinamento, e a partir dessas estimativas, a aplicação algoritmos otimizadores que identifiquem quais variáveis apresentam maior potencial de aprimoramento, permitindo a formulação de recomendações individuais. Dessa maneira, torna-se possível não apenas antecipar resultados, mas também propor intervenções direcionadas que contribuam para a maximização do desempenho esportivo.

Como objetivo secundário, propõe-se a integração desse sistema a um dispositivo capaz de operacionalizar as recomendações geradas em tempo hábil e de forma acessível ao usuário final. A intenção é que as informações produzidas pela modelagem sejam traduzidas em orientações claras, acionáveis e individualizadas, voltadas especialmente a atletas amadores, que frequentemente não dispõem de acompanhamento técnico especializado. Espera-se que essa abordagem contribua para aumentar o engajamento, sustentar a motivação ao longo do ciclo de treinamento e promover maior aderência às práticas recomendadas, reduzindo a probabilidade de abandono da corrida.

Esta dissertação posiciona-se na interseção entre a análise de perfis de corredores, a modelagem preditiva de desempenho em corrida e o desenvolvimento de recomendações baseadas em otimização. Embora a literatura apresente avanços relevantes nessas frentes de forma isolada, a integração sistemática entre clusterização, previsão de desempenho e geração de recomendações personalizadas permanece pouco explorada. O presente estudo contribui ao propor um fluxo analítico unificado que combina variáveis de caracterização individual obtidas por meio de pesquisa com informações detalhadas do histórico de treinamento, utilizando técnicas de clusterização para identificar perfis de corredores, modelos preditivos específicos por grupo para estimar o desempenho e, em um estágio subsequente, algoritmos de otimização para transformar essas previsões em recomendações práticas e acionáveis. Dessa forma, o trabalho avança além da identificação de fatores associados à performance, oferecendo um arcabouço orientado à ação, capaz de apoiar decisões individualizadas de treinamento com base em variáveis controláveis e nas particularidades de cada corredor.

Adicionalmente, a abordagem proposta reveste-se de particular relevância para atletas amadores, frequentemente desprovidos de acompanhamento profissional contínuo. Em um contexto marcado pela crescente popularidade da corrida e pela maior disponibilidade de dados relacionados a hábitos e histórico de treino, torna-se viável a oferta de análises rápidas e personalizadas, amparadas por técnicas de inteligência artificial, com vistas a subsidiar decisões estratégicas e otimizar o desempenho do atleta.

Este trabalho está organizado em capítulos que refletem uma progressão lógica. O Capítulo II reúne os principais conceitos, fundamentos metodológicos e definições utilizados ao longo do estudo, estabelecendo a base necessária para a compreensão das etapas subsequentes. O Capítulo III discute pesquisas prévias que guardam sinergia com esta investigação, destacando contribuições, limitações e oportunidades identificadas na literatura. No Capítulo IV, descrevem-se detalhadamente o desenho

da pesquisa, os procedimentos adotados para tratamento e modelagem dos dados, bem como a construção do fluxo de recomendação. O Capítulo V apresenta os achados obtidos, incluindo análises quantitativas, interpretações relevantes e *insights* derivados da aplicação do modelo proposto. Por fim, o Capítulo VI discorre sobre as limitações do estudo, possíveis aprimoramentos metodológicos, direções para trabalhos futuros, potenciais aplicações práticas da abordagem desenvolvida e as conclusões centrais alcançadas ao término da investigação.

Capítulo II Referencial Teórico

Este capítulo reúne os fundamentos conceituais e metodológicos que sustentam o desenvolvimento do estudo. São apresentados os principais conceitos, definições e pressupostos teóricos que norteiam a abordagem adotada, bem como os modelos e princípios que embasam as análises realizadas. O objetivo é consolidar o arcabouço teórico necessário para contextualizar as técnicas empregadas e dar suporte às etapas de modelagem, interpretação e avaliação desenvolvidas ao longo do trabalho.

A Seção II.1 apresenta os principais conceitos e técnicas utilizados para análise de dados ao longo do tempo, destacando métodos relevantes para identificação de padrões, tendências e sazonalidades. Em seguida, a Seção II.2 traz abordagens de aprendizado não supervisionado empregadas para segmentação e identificação de perfis semelhantes nos dados. A Seção II.3 descreve os fundamentos e modelos utilizados para estimar resultados futuros com base nas variáveis observadas, enfatizando as escolhas metodológicas adotadas no estudo. Por fim, a Seção II.4 discute os princípios e métodos aplicados para a busca de soluções ótimas, evidenciando seu papel no aprimoramento das recomendações e decisões derivadas das análises.

II.1 Conceitos de Séries Temporais

Formalmente, podemos considerar uma série temporal como uma sequência de vários pontos de observações indexados no tempo, ou seja, uma sequência de observações de um objeto de interesse coletadas ao longo do tempo. Quando as observações estão relacionadas a uma única variável, a série temporal é referenciada como univariada, caso contrário, quando as observações estão relacionadas a duas ou mais variáveis, a série pode ser considerada como multivariada [Ogasawara et al., 2025].

No âmbito de análise, a detecção de eventos em uma série temporal engloba a identificação de eventos passados (*offline*), a descoberta de eventos em tempo real (*online*) e até mesmo a previsão de eventos futuros, sendo considerada como uma função básica em sistemas de vigilância e monitoramento. Os conceitos são analisados de acordo com as especializações do tipo de evento, as estratégias de detecção de eventos abordadas e as métricas de avaliação. Geralmente, os eventos podem ser observados em anomalias, pontos de mudança ou *motifs* (padrões frequentes) [Ogasawara et al., 2025].

No contexto da previsão de series temporais, a técnica de janelas deslizantes - formalizada pela reconstrução do espaço de estados via vetores de *embedding* - transforma a estrutura temporal dos dados em um problema de aprendizado de maquina supervisionado, utilizando uma sequencia de n valores passados consecutivos $\{y_t, y_{t-1}, \ldots, y_{t-n+1}\}$ para prever o valor futuro $\{y_{t+1}\}$ [Bontempi et al., 2013]. Segundo Bontempi et al. [2013], essa abordagem permite capturar dependências temporais complexas e não lineares sem impor pressupostos rígidos sobre a dinâmica subjacente da serie, além de facilitar a adaptação continua do modelo a novos dados, sendo particularmente vantajosa quando integrada a métodos de aprendizado local, os quais evitam retreinamento oneroso e melhoram a robustez em cenários de não estacionariedade.

II.2 Algoritmos de Clusterização

Segundo Omran et al. [2007], a clusterização pode ser entendida como um processo de aprendizagem não supervisionada que busca identificar estruturas latentes em dados multidimensionais por meio da maximização da similaridade intra-grupo e da dissimilaridade inter-grupos. Esse processo envolve a definição de um espaço de atributos, a escolha de uma métrica de proximidade e a aplicação de algoritmos que induzem partições coerentes com a geometria dos dados. Entre as principais famílias de métodos, destacam-se as abordagens hierárquicas, que constroem dendrogramas a partir de fusões ou divisões sucessivas, e as particionais, como o *K-means* e suas variações (*Fuzzy C-means*, *K-harmonic means*, EM para misturas gaussianas), que formulam o problema como otimização iterativa da coerência interna dos agrupamentos. Cada escolha metodológica traz implicações, desde a simplicidade e sensibilidade à inicialização em algoritmos básicos até a maior flexibilidade e complexidade computacional de modelos mais sofisticados. Ainda de acordo com Omran et al. [2007], a avaliação da qualidade dos agrupamentos constitui uma dimensão fundamental do processo, sendo realizada por índices de validade que mensuram simultaneamente a compacidade *intra-cluster* e a separação entre *clusters*.

II.2.1 *K-means*

O algoritmo *K-means*, proposto originalmente entre as décadas de 1950 e 1960 por Steinhaus, Lloyd, MacQueen e Jancey, constitui uma das abordagens mais difundidas para a clusterização particional de dados [Ikotun et al., 2023]. Seu funcionamento baseia-se na minimização da soma dos erros quadráticos entre os pontos e o centróide do respectivo grupo, buscando obter a configuração que maximize a homogeneidade *intra-cluster* e a heterogeneidade *inter-cluster*. O processo é iterativo e envolve três etapas principais: a definição de k centróides iniciais, a atribuição de cada ponto ao centróide mais próximo (com base na métrica de distância Euclidiana) e a atualização dos centróides até que a partição se estabilize. Pela sua simplicidade de implementação, eficiência

computacional e boa escalabilidade, o *K-means* se consolidou como um dos algoritmos mais empregados em mineração de dados e aprendizado não supervisionado, figurando entre os dez métodos mais utilizados em aplicações de agrupamento [Ikotun et al., 2023].

Apesar dessas vantagens, o *K-means* apresenta limitações conhecidas, como a dependência da escolha inicial dos centróides e da definição prévia do número de *clusters* (k), o que pode levar a soluções locais subótimas. O uso da distância Euclidiana restringe sua capacidade de identificar agrupamentos não esféricos, enquanto a sensibilidade a *outliers* e dados ruidosos pode comprometer a qualidade dos resultados. Para mitigar essas limitações, diversas variantes e aprimoramentos têm sido propostos, como o *K-means++*, que melhora a inicialização dos centróides, o *X-means* e o *G-means*, que estimam automaticamente o número de *clusters*, além de abordagens adaptadas para grandes volumes de dados, como o *mini-batch K-means*. Tais evoluções refletem o contínuo interesse da comunidade científica em tornar o algoritmo mais robusto, flexível e aplicável a contextos complexos e de alta dimensionalidade [Ikotun et al., 2023].

II.2.2 *Aglomerative Hierarchical Clustering*

Enquanto o *K-means* opera a partir de uma partição pré-definida dos dados, as abordagens hierárquicas aglomerativas constroem a estrutura de agrupamentos de forma incremental, sem a necessidade de especificar o número de *clusters* antecipadamente. Conforme destacado por Tokuda et al. [2022], esses métodos procedem de modo *bottom-up*, iniciando com cada ponto como um *cluster* individual e sucessivamente fundindo os pares mais próximos de acordo com um critério de ligação. Entre os critérios clássicos estão as ligações única, completa, média, centróide, mediana e a de *Ward*, que minimiza a variância *intra-cluster*.

Uma vantagem central desses métodos é a capacidade de representar os agrupamentos por meio de dendrogramas, os quais permitem visualizar a hierarquia de fusões e selecionar o nível de corte mais adequado aos objetivos da análise. No entanto, a escolha do critério de ligação influencia significativamente a qualidade e a robustez dos *clusters* identificados. Estudos recentes, como o de Tokuda et al. [2022], mostram que métodos como a ligação única tendem a ser mais resilientes a falsos positivos em dados unimodais, enquanto abordagens como a de *Ward* podem detectar *clusters* espúrios com maior frequência. Essa família de métodos é especialmente útil quando se busca compreender a estrutura hierárquica dos dados ou quando não há conhecimento prévio sobre o número de *clusters*.

II.2.3 DBSCAN

O algoritmo DBSCAN também constitui uma das abordagens relevantes no campo das técnicas de agrupamento. O método parte do princípio de que um *cluster* é uma região do espaço de dados

caracterizada por alta densidade de objetos, separada por áreas de baixa densidade. A partir de dois parâmetros principais (o raio de vizinhança e o número mínimo de pontos) o algoritmo identifica três tipos de elementos: pontos centrais, que possuem densidade local suficiente; pontos de borda, que se situam próximos a regiões densas, mas não atingem o limiar mínimo de vizinhança; e pontos de ruído, que não pertencem a nenhum agrupamento. Essa lógica permite a detecção de aglomerados de formas arbitrárias, além da identificação automática de *outliers*, e também sem necessidade de conhecimento prévio sobre o número de grupos existentes no conjunto de dados [Bushra and Yi, 2021].

De acordo Bushra and Yi [2021], o DBSCAN apresenta como principais vantagens a robustez frente a ruídos, a capacidade de identificar *clusters* não lineares e a independência de suposições sobre a distribuição dos dados. Tais características o tornam aplicável em contextos diversos, no entanto, o desempenho do algoritmo está diretamente condicionado à definição adequada dos parâmetros iniciais. Valores muito baixos para o raio de vizinhança podem levar à fragmentação de agrupamentos, enquanto valores elevados tendem a unir grupos distintos em um mesmo *cluster*. Essa dependência acentuada dos parâmetros implica limitações na detecção de regiões com densidades distintas, problema recorrente em bases de dados reais e de alta complexidade [Bushra and Yi, 2021].

II.3 Modelagem Preditiva

II.3.1 Regressão Linear

A Regressão Linear é um dos métodos estatísticos e de aprendizado de máquina mais utilizados para modelagem e previsão, sendo aplicada em diversos campos devido à sua simplicidade e eficácia. O objetivo central dessa técnica é estabelecer a relação entre uma variável dependente e uma ou mais variáveis independentes, permitindo tanto previsões quanto análises de causalidade [Maulud and Abdulazeez, 2020]. O modelo pode assumir a forma de Regressão Linear Simples, quando há apenas um preditor, ou Regressão Linear Múltipla, quando múltiplas variáveis independentes são consideradas. Estudos recentes demonstram que, apesar de sua aparente simplicidade, a Regressão Linear continua sendo uma ferramenta poderosa na análise de dados, alcançando bons níveis de precisão em aplicações que vão desde previsões econômicas até diagnósticos médicos [Maulud and Abdulazeez, 2020].

Em termos mais técnicos, seu propósito é explicar a variabilidade de uma variável dependente Y por meio de uma ou mais variáveis independentes $X_1, X_2, \ldots, X_p$, estimando uma função linear que descreve essa relação [Olive, 2017]. O modelo de Regressão Linear Simples pode ser expresso pela equação

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{II.1}$$

em que β_0 representa o intercepto (valor esperado de Y quando $X = 0$), β_1 é o coeficiente angular (variação média de Y para uma unidade de variação em X) e ε_i é o termo de erro aleatório associado à observação i. Quando há mais de uma variável independente, obtém-se um modelo de Regressão Linear Múltipla.

A estimação dos parâmetros do modelo é tradicionalmente feita pelo método dos mínimos quadrados ordinários (*Ordinary Least Squares* (OLS)), que busca minimizar a soma dos quadrados dos resíduos. Esse estimador possui propriedades desejáveis sob determinadas condições, conhecidas como pressupostos clássicos do modelo linear, como linearidade, independência dos erros, homoscedasticidade, normalidade dos erros e ausência de multicolinearidade perfeita [Olive, 2017].

Quando esses pressupostos são atendidos, o estimador OLS é *Best Linear Unbiased Estimator* (BLUE), conforme o teorema de Gauss–Markov, ou seja, é o melhor estimador linear não viesado, com variância mínima entre todos os estimadores lineares não viesados. A qualidade do ajuste do modelo é comumente avaliada por meio do Coeficiente de Determinação (R^2), que indica a proporção da variância de Y explicada pelas variáveis independentes. Olive [2017] observa, contudo, que valores elevados de R^2 não garantem causalidade nem adequação do modelo, sendo essencial a análise dos resíduos e a verificação de possíveis violações dos pressupostos.

Em consonância com as observações de Maulud and Abdulazeez [2020], Olive [2017] reforça que a Regressão Linear, apesar do surgimento de métodos mais sofisticados de aprendizado de máquina, permanece essencial tanto para análise explicativa quanto preditiva. Sua simplicidade, interpretabilidade e base teórica sólida fazem dela uma ferramenta de referência na análise de dados, servindo como ponto de partida para modelos mais complexos e como meio de validação de relações lineares em contextos empíricos.

II.3.2 XGBoost

O XGBoost é um método de aprendizado supervisionado baseado em *gradient boosting* com árvores de regressão, amplamente utilizado devido à sua elevada performance preditiva e eficiência computacional [Chen and Guestrin, 2016]. O modelo constrói um conjunto de árvores de forma aditiva, em que a predição final é obtida pela soma das saídas individuais de cada árvore:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \tag{II.2}$$

onde cada função f_k representa uma árvore de regressão, que associa cada observação a um valor contínuo em uma de suas folhas [Chen and Guestrin, 2016].

Um dos principais diferenciais do XGBoost é a formulação de uma função objetivo regularizada, que busca minimizar simultaneamente o erro de predição e a complexidade do modelo. A função objetivo é definida como:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{II.3}$$

em que $l(\cdot)$ é uma função de perda convexa e diferenciável — como o erro quadrático médio no caso de regressão — e $\Omega(f)$ é um termo de regularização que penaliza a complexidade das árvores. Esse termo é dado por:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{II.4}$$

onde T representa o número de folhas da árvore, w os valores preditivos associados às folhas, e γ e λ são hiperparâmetros de regularização [Chen and Guestrin, 2016].

O processo de treinamento ocorre de forma iterativa, adicionando uma nova árvore a cada etapa com o objetivo de corrigir os erros residuais do modelo anterior. Para tornar essa otimização eficiente, o XGBoost utiliza uma aproximação de segunda ordem da função de perda, baseada no gradiente e na hessiana, o que permite avaliar de forma mais precisa o ganho obtido por cada divisão da árvore [Chen and Guestrin, 2016]. Essa abordagem resulta em melhor convergência e maior estabilidade, especialmente em problemas de regressão com grandes volumes de dados.

Além disso, o XGBoost incorpora mecanismos como regularização explícita, *shrinkage* e tratamento nativo de dados esparsos. O *shrinkage*, implementado por meio de um fator de aprendizagem que reduz a contribuição de cada árvore adicionada ao modelo, permite que o processo de otimização ocorra de forma mais gradual. Esse mecanismo, em conjunto com a penalização da complexidade das árvores, contribui para a redução de *overfitting* e para o aumento da capacidade de generalização do modelo, além de favorecer a escalabilidade do algoritmo em cenários de alta dimensionalidade [Chen and Guestrin, 2016].

II.4 Algoritmos Otimizadores

Os algoritmos de otimização desempenham um papel central no desenvolvimento e aprimoramento de modelos de aprendizado de máquina, constituindo-se como ferramentas potentes para resolver problemas complexos que envolvem múltiplas variáveis, uma vez que a maioria dos algoritmos se baseia na minimização de funções de custo ou erro. De acordo com Kumar [2024], esses algoritmos de otimização englobam uma ampla gama de técnicas que vão desde a programação matemática clássica até métodos inspirados em fenômenos naturais, que permitem ajustar parâmetros e pesos de modelos preditivos, assegurando uma convergência mais rápida e estável para soluções

que generalizem adequadamente os padrões observados nos dados. Sua principal finalidade é encontrar soluções ótimas, ou quase ótimas, dentro de um espaço de busca, o que os torna fundamentais para a eficiência de aplicações em áreas como engenharia, economia, ciência de dados e inteligência artificial [Kumar, 2024].

Estratégias de otimização contemporâneas vêm se tornando cada vez mais relevantes em aplicações de larga escala, especialmente em ambientes de aprendizado dinâmico e distribuído, como redes neurais profundas, sistemas de recomendação e otimização de hiperparâmetros [Kumar, 2024]. A biblioteca SciPy[1] destaca-se como uma das principais ferramentas para a implementação de algoritmos de otimização em Python, oferecendo um amplo conjunto de rotinas numéricas aplicáveis a diferentes áreas. O módulo `scipy.optimize` fornece algoritmos fundamentais para solução de problemas de otimização linear e não linear, enraizados em métodos clássicos de cálculo numérico, e constitui um dos pilares da biblioteca, que é amplamente utilizada em pesquisa e aplicações industriais.

A função `scipy.optimize.minimize` é descrito por Virtanen et al. [2020] como uma interface unificada para encontrar mínimos locais de funções não lineares. Ele reúne diferentes métodos, como Nelder-Mead, *Broyden–Fletcher–Goldfarb–Shanno* (BFGS), *Limited-memory Broyden–Fletcher–Goldfarb–Shanno - Bounded* (L-BFGS-B), *Sequential Least Squares Programming* (SLSQP), *Constrained Optimization by Linear Approximations* (COBYLA), Trust-Region e Newton-CG, entre outros, permitindo ao pesquisador selecionar o mais adequado conforme a natureza do problema. Seu princípio básico consiste em iterar sobre um espaço de parâmetros em busca do ponto que reduz ao máximo o valor da função objetivo, utilizando gradientes e, quando disponíveis, derivadas de segunda ordem para acelerar a convergência.

Os parâmetros de inicialização têm papel relevante na qualidade e velocidade da solução. O vetor inicial fornece o ponto de partida para o processo iterativo, influenciando a probabilidade de convergir a mínimos locais ou globais. O `minimize` também permite especificar limites (*bounds*) para as variáveis, garantindo que as soluções obtidas respeitem restrições de domínio físico ou lógico do problema [Virtanen et al., 2020]. Essa funcionalidade é particularmente útil em modelos de otimização de hiperparâmetros ou em contextos aplicados, como alocação de recursos ou necessidade de se encontrar melhores valores para variáveis independentes. Outro aspecto relevante é o tratamento das restrições (*constraints*). A função suporta tanto restrições de igualdade quanto de desigualdade, expressas como funções das variáveis de decisão. Isso amplia a aplicabilidade do método para problemas reais que exigem que certas condições sejam mantidas ao longo da otimização, por exemplo, manter proporções, somatórios fixos ou limites operacionais.

Conforme detalhado por Virtanen et al. [2020], o `minimize` implementa métodos que lidam di-

[1]Disponível em: https://scipy.org/pt/. Último acesso em: 20 de outubro de 2025.

retamente com restrições não lineares, como SLSQP, que será explorado no presente estudo. O método é amplamente utilizado na resolução de problemas de otimização com restrições, utilizando uma abordagem baseada em programação quadrática sequencial, que resolve iterativamente subproblemas de aproximação quadrática, sendo adequado para restrições de igualdade e desigualdade diferenciáveis [Kraft, 1988], da forma:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{sujeito a} \quad & g_i(\mathbf{x}) = 0, \\ & h_j(\mathbf{x}) \geq 0, \\ & \mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \end{aligned} \tag{II.5}$$

Em cada iteração k, o problema original é aproximado por um subproblema quadrático, no qual a função objetivo é linearizada e regularizada por uma aproximação quase-Newton da Hessiana:

$$\min_{\mathbf{d}} \ \nabla f(\mathbf{x}_k)^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \mathbf{B}_k \mathbf{d} \tag{II.6}$$

sujeito às restrições linearizadas no ponto $\mathbf{x}_k$. A direção de busca $\mathbf{d}$ é então utilizada para atualizar a solução:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}, \tag{II.7}$$

onde α_k é determinado por uma busca linear que garante a redução da função objetivo e a viabilidade das restrições [Kraft, 1988].

Por fim, embora o `minimize` seja projetado para minimizar funções, ele também pode ser usado para maximizar objetivos: esse processo pode ser realizado de maneira simples, no qual basta inverter o sinal da função (multiplicando-a por -1). Essa estratégia é comum em aprendizado de máquina e em modelagem, quando o objetivo é maximizar a acurácia, o lucro ou outro indicador de desempenho.

Capítulo III Trabalhos Relacionados

Este capítulo apresenta uma revisão de pesquisas que se relacionam diretamente com os objetivos e a abordagem desta investigação. São discutidos estudos que exploram métodos de previsão de desempenho, técnicas de recomendação, modelagem aplicada ao contexto esportivo e análises de desempenho segmentadas por perfil de atleta. A revisão contempla também trabalhos que empregam estratégias de otimização em cenários semelhantes. A análise reúne as principais contribuições encontradas na literatura, bem como limitações recorrentes e oportunidades de avanço, que motivam e justificam o desenvolvimento da proposta aqui delineada.

A Seção III.1 reúne estudos que aplicam métodos de análise temporal para monitoramento e previsão de métricas de desempenho esportivo, evidenciando suas principais contribuições. A Seção III.2 examina pesquisas que utilizam técnicas de segmentação para identificar perfis de atletas e padrões de comportamento, destacando abordagens metodológicas recorrentes na literatura. Em seguida, a Seção III.3 apresenta investigações focadas em modelos preditivos aplicados a métricas de corrida, discutindo variáveis consideradas relevantes e estratégias de modelagem empregadas. A Seção III.4 compila estudos que utilizam métodos de otimização para apoiar a tomada de decisão e melhorar resultados em diferentes contextos. Posteriormente, a Seção III.5 aborda pesquisas voltadas à geração de recomendações individualizadas com base em dados de usuários ou atletas. Por fim, a Seção III.6 sintetiza as lacunas identificadas na literatura revisada e explicita a contribuição específica desta pesquisa frente aos trabalhos existentes.

III.1 Séries Temporais no Esporte

A utilização de séries temporais tem se consolidado como ferramenta relevante para a análise esportiva. Em sua revisão, Komitova et al. [2022] destacam aplicações que vão desde a previsão de desempenho e detecção de padrões de carga de treino até a identificação de eventos críticos, como risco de lesão e fadiga. O estudo evidencia que técnicas de mineração de séries temporais oferecem suporte valioso para monitoramento contínuo de atletas, auxiliando tanto em estratégias de otimização de performance quanto na prevenção de problemas de saúde relacionados ao esporte.

Corbett et al. [2019]; Teune et al. [2022] aplicaram técnicas de séries temporais com enfoque em *change point analysis* para compreender o desempenho de atletas em esportes coletivos, parti-

cularmente no futebol australiano. Esses estudos demonstraram como a identificação de pontos de mudança ao longo das partidas ou atividades de treino permite captar transições importantes no comportamento físico e técnico dos jogadores, como variações de intensidade, eficiência ou padrões de movimentação. Essa abordagem possibilita não apenas descrever a dinâmica do jogo em tempo real, mas também oferecer subsídios práticos para o desenho de treinamentos, por exemplo, ao indicar durações ideais de atividades, momentos de queda de rendimento ou ajustes de carga. Assim, a análise de pontos de mudança contribui para aproximar dados de desempenho da prática esportiva, permitindo decisões mais embasadas para otimizar tanto a performance quanto a prevenção de fadiga e lesões.

Na corrida, Nakayama et al. [2010] utilizaram análise de séries temporais pra comparar corredores treinados com não corredores e descobriram que os corredores apresentam um padrão de passada mais consistente e menos variável. Usando métodos avançados de análise, os autores mostraram que os corredores experientes não só correm em um ritmo preferido mais rápido, mas também mantêm uma cadência de passada mais estável, mesmo em diferentes velocidades. Esse achado sugere que o treinamento contínuo na corrida permite um movimento mais eficiente e controlado, refletindo uma melhor adaptação neuromuscular e um maior aproveitamento da experiência adquirida.

Ainda nessa temática, se apoiando no paradigma de aprendizagem por reforço, Van den Berghe et al. [2020] investigaram como corredores reagiram a um sistema interativo de *feedback* musical em tempo real, projetado para reduzir o impacto da corrida. Os autores utilizaram detecção de eventos em séries temporais para detectar o momento exato em que os corredores adaptaram sua técnica e reduziram a aceleração tibial de pico. Os resultados mostraram que os corredores foram capazes de responder ao *feedback* auditivo de forma rápida e individualizada, alcançando uma redução significativa no impacto e destacando o potencial do uso de *feedback* na modificação da técnica de corrida em ambiente natural.

III.2 Análise de *Cluster* no Esporte

A análise de *cluster* tem sido amplamente aplicada na corrida para identificar diferentes perfis de atletas. Para compreender a diversidade entre corredores não profissionais, Thuany et al. [2020] utilizaram Análise de Classe Latente em uma amostra nacional de corredores brasileiros, identificando dois perfis distintos: “amadores” e “recreativos”. As diferenças entre as classes foram explicadas principalmente por variáveis de treinamento (volume e frequência semanal), Índice de massa corporal (IMC) e motivação para a prática. Enquanto os recreativos tendiam a valorizar mais saúde e qualidade de vida, com menor engajamento em treinos, os amadores mostraram maior dedicação, motivação voltada ao desempenho e melhor ritmo de corrida. Os autores destacam ainda que homens, atletas mais jovens e classificados como amadores apresentaram maior probabilidade

de alcançar melhor performance, reforçando a relevância do perfil do corredor na compreensão do desempenho esportivo.

O estudo de Wang et al. [2021] também utilizou análise de agrupamento para caracterizar padrões de corrida, identificando dois perfis principais: corredores “sustentados”, que mantêm ritmo consistente ao longo do tempo, e “menos sustentados”, que apresentam maior variação no desempenho. A partir de dados coletados via dispositivos móveis e informações geográficas, os autores exploraram como condições temporais (horário, dia da semana) e ambientais (clima, poluição do ar, densidade populacional) impactam a performance. Os resultados mostraram que esses fatores não afetam todos os corredores da mesma forma: enquanto os do perfil sustentado tendem a apresentar menor sensibilidade a variações externas, os menos sustentados são mais influenciados pelo contexto, sugerindo que aspectos ambientais devem ser considerados na personalização de treinos e estratégias de corrida.

Uma análise de agrupamento também foi usada por Rodriguez Rivadulla et al. [2024] para identificar grupos de corredores com técnica semelhante com base diversas variáveis cinemáticas que descrevem a técnica de corrida. Assim como nos outros trabalhos mencionados, também foram encontrados dois grupos de corredores que mostraram cinemática distinta: o grupo de "pelve neutra"e o grupo de "pelve inclinada", no qual analisa como esses perfis se adaptam à velocidade de forma diferente.

Além de identificar perfis de corredores, a análise de agrupamento também pode auxiliar na compreensão de padrões de desempenho relacionados a lesões. Jauhiainen et al. [2020] aplicaram a análise hierárquica de *clusters* em dados de corredores saudáveis e lesionados, com o objetivo de verificar se existiam subgrupos homogêneos associados a diferentes tipos de lesões. O estudo identificou padrões distintos de corrida, mas observou que corredores com lesões semelhantes não se agruparam necessariamente nos mesmos *clusters*. Esses achados sugerem que padrões biomecânicos de corrida existem independentemente da localização da lesão, ressaltando a importância de considerar a variabilidade dos movimentos ao planejar estratégias de prevenção e reabilitação.

Apesar dos avanços proporcionados pela literatura, observa-se que a maior parte dos estudos emprega a clusterização com finalidade predominantemente descritiva, concentrando-se na identificação e caracterização de perfis de corredores. Essa abordagem, embora relevante para compreender a heterogeneidade entre atletas, limita o potencial de uso dos agrupamentos em etapas analíticas subsequentes. Nesse sentido, emerge uma oportunidade de investigação na integração entre a segmentação por *clusters* e o desenvolvimento de modelos preditivos específicos por perfil, ampliando o papel da clusterização para além da descrição e explorando seu potencial como componente estruturante em pipelines analíticos mais completos.

III.3 Previsão de Desempenho na Corrida

A predição de desempenho é uma ferramenta valiosa para atletas e treinadores. Em estudo recente com atletas de alto rendimento, Wiecha et al. [2022] propuseram uma abordagem de aprendizado de máquina para prever velocidades de corrida em diferentes intensidades fisiológicas. Os autores utilizaram primeiramente um algoritmo baseado em árvore (XGBoost) para selecionar os preditores mais relevantes na previsão de desempenho e em seguida, aplicaram modelos de Regressão Linear Múltipla separadamente para homens e mulheres, a fim de estimar a velocidade no limiar anaeróbico no Ponto de Compensação Respiratória (RCP) e a velocidade máxima. O estudo, conduzido com uma ampla amostra de atletas, demonstrou alta precisão preditiva e os resultados destacam que a capacidade aeróbica (VO_2) e a eficiência no manejo metabólico do lactato em esforços submaximais foram os determinantes primários do desempenho de corrida, reforçando a utilidade desses modelos para a prescrição personalizada de treino [Wiecha et al., 2022].

Com a crescente popularização e participação de atletas amadores em competições de corrida, a demanda por ferramentas de avaliação e prescrição de treino expande-se além do âmbito de alto rendimento. Nesse contexto, Knechtle et al. [2011] investigaram variáveis preditoras para o tempo de prova em meia maratona (21 km) em corredoras recreacionais. Por meio de um modelo de Regressão Linear Múltipla, os autores identificaram que a espessura da dobra cutânea da axila média e a velocidade média dos treinos eram os fatores mais fortemente associados ao desempenho na corrida. Ou seja, para essa população, indicadores antropométricos simples e métricas de treino podem ser utilizados para estimar o desempenho em distâncias de resistência com considerável precisão.

Seguindo mesma direção, Shu et al. [2024] propuseram um modelo de previsão do desempenho em meia maratona (21 km) especificamente para corredores recreacionais do sexo masculino. Foram consideradas variáveis de diferentes dimensões, incluindo características de treinamento (anos de prática, frequência semanal, distância mais comum dos treinos, padrão de pisada), fatores de saúde (doenças respiratórias, hipertensão, lesões prévias, dependência ao exercício), além do uso de suplementos (como cafeína, creatina e *whey protein*). Embora muitas dessas variáveis não tenham se mostrado preditoras significativas, a análise multivariada destacou três fatores centrais associados a melhores desempenhos: maior distância mensal percorrida, ritmo médio de treino mais rápido e melhor qualidade do sono. A partir desses indicadores, foi desenvolvido um nomograma com bom poder discriminativo, capaz de estimar a probabilidade de um atleta completar a prova em até 105 minutos [Shu et al., 2024]. Esses resultados reforçam que, além das métricas de treino, aspectos relacionados à recuperação e saúde desempenham papel crucial na previsão de performance em corridas de longa distância.

Ainda corroborando essa perspectiva, Suwankan et al. [2024] investigaram se características físicas e de treinamento estariam associadas ao desempenho de corredores recreativos em uma prova de 10 km. O estudo envolveu 60 corredores e analisou variáveis físicas e de treinamento (dias, distância e tempo semanais, experiência em corrida e participação em provas). Os resultados mostraram que o tempo de prova se correlacionou significativamente com o físico dos atletas, frequência e volume de treino, experiência prévia e número de corridas disputadas, sendo que a distância semanal de treino apresentou o maior impacto no desempenho. De modo geral, os achados indicam que tanto aspectos físicos quanto a consistência e o volume de treinamento são determinantes para a inferir o desempenho em corredores recreativos.

No contexto de séries temporais de treinamento, El-Kassabi et al. [2020] aplicaram modelos de *Deep Learning* para prever o desempenho de atletas em maratonas, utilizando dados públicos de treinos da *Boston Marathon*. Foram avaliadas arquiteturas como *Recurrent Neural Network* (RNN), *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU), explorando variáveis como distância, duração, ritmo e ganho de elevação. Os resultados mostraram que os modelos de *Deep Learning* superaram abordagens tradicionais, como a Regressão Linear, oferecendo maior precisão na previsão do tempo final de prova.

Além de fatores intrínsecos ao atleta, como características e condicionamento físico, parâmetros ambientais também podem exercer influência significativa no desempenho em corridas. Um estudo de El Helou et al. [2012] analisou o impacto de variáveis climáticas e de poluição atmosférica em seis maratonas internacionais entre 2001 e 2010, utilizando uma amostra de mais de 1,7 milhão de participantes. Os resultados demonstraram que a temperatura do ar é o fator ambiental de maior relevância, correlacionando-se com a velocidade de corrida: para cada nível de desempenho, existe uma temperatura ótima específica, e desvios acima ou abaixo desse ponto resultam em redução de desempenho e aumento na taxa de desistência. Ainda de acordo com o estudo, a umidade também apresentou correlação significativa, embora menor, sugerindo que o processo de previsão de desempenho em maratonas deve considerar tais variáveis ambientais para uma avaliação mais precisa e contextualizada.

Na Tabela III.1 são apresentados relevantes estudos que aplicaram modelos de previsão de desempenho na corrida, indicando os autores e os algoritmos utilizados em cada trabalho. Esta síntese permite visualizar as abordagens mais comuns e identificar padrões ou lacunas na aplicação de técnicas de previsão no contexto esportivo.

Tabela III.1: Estudos de previsão de desempenho na corrida

Autores	**Atributos**	**Algoritmos**
Rebbouj and Lotfi [2024]	Idade, peso, altura, IMC, força muscular, frequência cardíaca e tempo em 12 min.	Regressão Linear
Shu et al. [2024]	Idade, IMC, estado civil, filhos, escolaridade, ocupação, gasto com equipamento, treino (volume e experiência), tempo de aquecimento, padrão de pisada, lesões, histórico de doenças, qualidade do sono, vício em exercício e suplementos usados.	Regressão Logística, Nomograma, Regressão Linear
Suwankan et al. [2024]	IMC, dias de treino/semana, distância de treino/semana, duração do treino/semana, experiência em anos e número de participações em corridas.	Regressão Linear
Johansson et al. [2024]	Tempos parciais (5 km, 10 km, etc.), idade, sexo, temperatura e histórico de corridas.	Regressão Linear, Rede Neural, Random Forest
Thuany et al. [2023]	Idade, sexo, tempo de prova, ano do evento e ritmo de corrida.	Regressão Linear
Lerebourg et al. [2023]	Tempo nos 10 km, IMC, idade e sexo.	*K-Nearest Neighbors* (KNN), *Artificial Neural Network* (ANN)

Tabela III.1 (continuação)

Autores	**Atributos**	**Algoritmos**
Wiecha et al. [2022]	Idade, IMC, percentual de gordura e variáveis fisiológicas como Consumo máximo de oxigênio (VO_2max), Consumo de oxigênio no ponto de compensação respiratória (VO_2RCP), Consumo de oxigênio no limiar anaeróbico (VO_2AT), Concentração de lactato no sangue no ponto de compensação respiratória (La^-bRCP), Concentração de lactato no sangue no limiar anaeróbico (La^-bAT), Ventilação pulmonar máxima (VEmax), Ventilação pulmonar no ponto de compensação respiratória (VERCP).	XGBoost e Regressão Linear
El-Kassabi et al. [2020]	Distância, tempo, ritmo, elevação e dados de treino sequenciais.	RNN, LSTM, GRU, Regressão Linear
Yong et al. [2020]	Tempo de chegada, idade, gênero, ano da maratona, média e desvio padrão dos tempos.	*Autoregressive Moving Average* (ARMA)
Gómez-Molina et al. [2017]	Antropometria (altura, massa, IMC, dobras), treino (volume, experiência), fisiologia (VO_2max, velocidade no RCP) e biomecânica (comprimento de passada, cadência).	Regressão Linear

Tabela III.1 (continuação)

Autores	**Atributos**	**Algoritmos**
Muñoz et al. [2012]	Velocidade em teste de pista e concentração de lactato sanguíneo).	Regressão Linear
Tanda [2011]	Distância média semanal e ritmo médio de treino.	Regressão Linear e Exponencial
Knechtle et al. [2011]	Espessura de dobras cutâneas (axilar média, subescapular, etc.), IMC, percentual de gordura e velocidade média dos treinos.	Regressão Linear
Billat et al. [2001]	VO_2max, custo energético, frequência cardíaca, concentração de lactato sanguíneo, ritmo em 1000 m e volume de treino.	Regressão Linear

De modo geral, os estudos revisados apresentam contribuições importantes para a predição de desempenho em corrida, porém concentram-se, em sua maioria, em populações específicas, como atletas de alto rendimento, corredores recreacionais de determinado sexo ou amostras com características bastante delimitadas, o que pode restringir a generalização dos achados para contextos mais heterogêneos. Além disso, observa-se que grande parte das abordagens privilegia modelos globais, sem explorar de forma sistemática possíveis diferenças estruturais entre perfis de corredores. Nesse cenário, há uma oportunidade de investigação em abordagens capazes de conciliar a predição em nível geral com análises segmentadas por grupos homogêneos de atletas, potencialmente ampliando a robustez e a aplicabilidade dos modelos de desempenho.

III.4 Otimização de Variáveis Para Maximizar um Objetivo

No contexto de alavancar modelos preditivos para decisões prescritivas, Mišić [2020] propõe uma abordagem para otimização de conjuntos de árvores. O autor investiga o problema de determinar a configuração ótima de variáveis independentes que sejam controláveis e que possam maximizar a saída prevista por um modelo de ensemble, como florestas aleatórias ou modelos impulsionados,

quando este é empregado como uma função objetivo. A contribuição central do trabalho reside na formulação do problema como um modelo de Otimização Inteira Mista (MIO), que representa explicitamente a estrutura lógica de cada árvore no conjunto. Mišić [2020] desenvolve e analisa teoricamente uma hierarquia de formulações aproximadas baseadas em profundidade e a eficácia metodológica é validada empiricamente em estudos de caso de *design* de medicamentos e precificação personalizada, onde a abordagem mostrou-se capaz de alcançar o ponto ótimo.

De maneira análoga, no estudo de Cohen et al. [2017] os autores abordaram um problema de otimização de promoções no varejo de bens de consumo. Eles desenvolveram um modelo de otimização baseado em programação linear inteira para selecionar preços promoções ao longo de um horizonte de tempo, com o objetivo de maximizar o lucro do varejista. O modelo incorpora regras de negócio práticas, como a limitação do número de promoções e a imposição de períodos de separação entre promoções consecutivas, além de considerar efeitos de pós-promoção na demanda, como a redução nas vendas após uma promoção. A abordagem foi validada em dados reais e permitiu um aumento de lucro, demonstrando o valor da otimização de variáveis controláveis, como preços e cronogramas de promoção, para a maximização de desempenho operacional e financeiro.

Avançando na aplicação de modelos de otimização, Kakaew and U-tapao [2023] propuseram um modelo de programação linear para maximizar o lucro em empreendimentos habitacionais na região de Bangkok e arredores, na Tailândia. Foi utilizada uma equação de regressão, obtida a partir de dados históricos de custos de terra e construção, como função objetivo em um modelo de otimização. Considerando restrições orçamentárias e de área, os autores otimizaram variáveis controláveis, como a quantidade de unidades de cada tipo de habitação a serem construídas, permitindo identificar a combinação de tipologias que maximiza o lucro total do empreendimento. Para complementar a análise, foi realizada ainda uma análise de sensibilidade, explorando cenários alternativos e reforçando a utilidade da abordagem como ferramenta de apoio ao planejamento de projetos em cidades densamente povoadas.

III.5 Recomendações Personalizadas

No campo do suporte ao desempenho esportivo, Smyth [2019] destaca a relevância dos sistemas de recomendação aplicados a esportes como corrida e triatlo, para oferecer programas de treino mais personalizados. A partir do histórico de treinamento dos atletas, contemplando variáveis como carga, recuperação, tipo de sessão e resultados anteriores, esses sistemas permitem adaptar recomendações de forma individualizada, auxiliando na escolha do volume, intensidade e estratégias de preparação. Essa personalização é apontada como um fator crucial para otimizar o desempenho, ao alinhar o planejamento de treinos às necessidades e características específicas de cada atleta.

Complementarmente à personalização dos planos de treino, a literatura recente explora o po-

tencial dos sistemas de recomendação para intervenções em tempo real durante a competição. No contexto de provas, Berndsen et al. [2019] propuseram um sistema que utiliza aprendizado de máquina e análise de séries temporais para gerar recomendações de ritmo enquanto o corredor está na prova. A partir de dados de ritmo, frequência cardíaca e cadência de corredores, o modelo identifica atletas em risco de desacelerar drasticamente recomendando assim um ritmo personalizado. Essa recomendação é baseada no comportamento de corredores similares que conseguiram evitar a desaceleração, oferecendo assim uma estratégia corretiva dinâmica e explicável, que visa otimizar o tempo final [Berndsen et al., 2019].

Ainda nesse contexto, Smyth et al. [2022] realizam uma contribuição relevante ao explorar de forma abrangente sistemas de recomendação em maratona. Os autores propõem uma visão estruturada de como técnicas de aprendizado de máquina e sistemas de recomendação podem ser utilizados para apoiar corredores recreacionais em todas as etapas de sua jornada, que muitas das vezes pode ser bem difícil e exigir grande capacidade psicológica. O estudo e organiza uma ampla gama de oportunidades de recomendação, classificando-as em tarefas primárias, como recomendação de planos de treino personalizados, estimativa de condicionamento físico, predição de risco de lesões e sugestão de estratégias de ritmo; e em tarefas secundárias, mais alinhadas ao comércio eletrônico tradicional, como recomendação de eventos, equipamentos, rotas e parceiros de treino.

Para validar a proposta, os autores apresentam quatro estudos de caso que implementam e avaliam algumas das tarefas primárias identificadas e que abordam problemas centrais no treinamento para maratonas. O primeiro estudo de caso, uma contribuição original, foca na estimativa de modelos de condicionamento físico personalizados a partir de dados brutos de treino, como séries de ritmo coletadas por aplicativos. Os autores demonstram que é possível inferir métricas fisiológicas cruciais, como o VO_2max e a velocidade crítica, sem a necessidade de testes laboratoriais dispendiosos. Ao utilizar os ritmos mais rápidos alcançados por um corredor em diferentes distâncias durante seu treinamento normal, o modelo combinado foi capaz de prever o tempo de maratona com um erro médio bem baixo [Smyth et al., 2022].

O segundo e o terceiro estudos de caso também consistem em pesquisas originais. O segundo investiga a recomendação personalizada de sessões de treino usando Raciocínio Baseado em Casos (CBR). A abordagem recupera a semana de treino de um corredor histórico, com perfil e desempenho similares, que alcançou um tempo de maratona meta. Os resultados mostram que as recomendações geradas são consistentes: planos para tempos mais rápidos sugerem aumentos no volume semanal e ritmos médios mais velozes, e vice-versa. Já o terceiro estudo de caso aborda o monitoramento proativo do risco de lesões. Diante da ausência de dados explícitos de lesão, os autores utilizam interrupções prolongadas no treinamento como um *proxy*. Embora a classificação binária (lesão/não-lesão) tenha tido precisão limitada, a probabilidade da classe positiva mostrou-se

altamente correlacionada com a ocorrência de interrupções, oferecendo um valioso "índice de risco de lesão"que pode alertar os corredores para modificar sua carga de treino [Smyth et al., 2022].

O quarto e último estudo de caso resume um trabalho prévio da equipe, aplicando o CBR ao planejamento da prova em si. O sistema utiliza o histórico de maratonas anteriores de um corredor para prever um tempo pessoal desafiador porém realizável e, crucialmente, para recomendar um plano de ritmo segmentado para a corrida-alvo. A estratégia baseada na média dos casos similares superou outras, gerando previsões de tempo com erro médio também baixo e planos de ritmo altamente similares aos adotados com sucesso por corredores de perfil análogo. Esse caso evidencia o papel do sistema de recomendação não apenas no treinamento, mas também na estratégia de execução da prova, fechando o ciclo de suporte ao atleta [Smyth et al., 2022].

É notório que a sistemas de recomendação têm ganhado crescente relevância no contexto esportivo, atuando como ferramentas estratégicas para otimizar o desempenho atlético por meio de análise automatizada de dados. Como destacado por V. et al. [2022], tais sistemas permitem uma avaliação contínua e personalizada do estado do atleta, integrando variáveis biológicas, ambientais e pessoais. Essa abordagem não apenas auxilia na melhoria da qualidade do treinamento, mas também contribui para a longevidade do atleta, reduzindo a dependência de intervenção humana direta e facilitando a tomada de decisão com base em dados armazenados em ambientes de nuvem.

No estudo em questão, os autores propõem uma abordagem híbrida que combina a Computação Evolutiva Multi-Gradiente Extrema (EMGEC) com a Otimização do Lobo Cinzento Aprimorada (IGWO). O algoritmo de IGWO atua na seleção de características mais relevantes, assegurando um equilíbrio entre exploração e exploração do espaço de busca, enquanto o EMGEC é empregado para otimização multiobjetivo, considerando métricas de desempenho da modelagom, como precisão, *recall* e novidade nas recomendações. A combinação dessas técnicas visa superar limitações de métodos tradicionais.

A incorporação de técnicas de otimização avançada, como o IGWO e o EMGEC, evidencia o potencial da inteligência computacional para lidar com problemas complexos e multiobjetivo no domínio esportivo [V. et al., 2022]. Conforme demonstrado pelos autores, a otimização não apenas melhora a acurácia do sistema de recomendação, mas também permite uma adaptação dinâmica às condições variáveis do atleta e do ambiente, resultando em um suporte mais robusto e personalizado ao longo do processo de treinamento.

Apesar dos avanços observados, a literatura tem se concentrado majoritariamente em sistemas de recomendação no esporte baseados em similaridade entre atletas ou na inferência indireta do condicionamento físico a partir de dados históricos. Embora eficazes para personalização, essas abordagens tendem a tratar a recomendação de forma implícita, com menor ênfase na otimização direta de variáveis de treino passíveis de controle. Nesse sentido, identifica-se uma oportunidade

de investigação em estratégias que formulem o problema de recomendação sob uma perspectiva explícita de otimização, considerando simultaneamente restrições operacionais e as particularidades estruturais de diferentes grupos de corredores, o que pode ampliar a precisão e a utilidade prática das recomendações geradas.

III.6 Posicionamento do Estudo

A presente dissertação situa-se na interseção entre análise de perfis de corredores por meio de técnicas de clusterização, modelagem preditiva de desempenho em corrida e mecanismos de recomendação baseados em otimização. Embora existam estudos consolidados nessas três vertentes, observa-se que grande parte das pesquisas concentra-se em abordagens isoladas, seja apenas na previsão de performance, na caracterização de perfis de corredores ou na proposição de sistemas de recomendação, enquanto a integração sistemática dessas dimensões permanece pouco explorada. Nesse contexto, o trabalho aqui desenvolvido contribui ao combinar esses três eixos em um único modelo analítico orientado à geração de recomendações práticas e personalizadas, fundamentadas tanto em variáveis pessoais quanto em características detalhadas do histórico de treinamento.

No âmbito da previsão de desempenho, estudos tradicionais como o de Wiecha et al. [2022], Knechtle et al. [2011], Shu et al. [2024] e Suwankan et al. [2024] demonstram que características físicas, métricas de treino e, em alguns casos, fatores comportamentais influenciam de modo significativo o tempo de prova em diferentes distâncias. Embora esses trabalhos ofereçam modelos robustos, eles se concentram majoritariamente em populações específicas (como mulheres recreacionais, atletas de alto rendimento ou amostras maiores e mais homogêneas) e adotam, em geral, modelos exclusivamente preditivos, sem integração direta com mecanismos de recomendação. O presente estudo avança na literatura ao desenvolver um modelo preditivo capaz de capturar esses determinantes, mas associado a um segundo estágio analítico (otimização) que transforma previsões em recomendações diretamente aplicáveis ao treinamento do corredor. Assim, diferentemente dos estudos anteriores que se restringem à identificação de variáveis preditoras, aqui as informações são operacionalizadas para orientar melhorias individualizadas, considerando variáveis controláveis como distância, frequência de treino e intensidade.

No que se refere à análise de perfis de corredores por clusterização, uma diferença central é que, enquanto os estudos citados buscam descrever perfis, o presente trabalho utiliza os *clusters* como base para modelagens preditivas específicas por grupo, permitindo estimativas mais ajustadas do desempenho e interpretação diferenciada das variáveis determinantes. Esse encadeamento metodológico que envolve `clusterização → modelos específicos por` *`cluster`* `→ otimização`, não aparece de forma integrada na literatura revisada, representando uma contribuição original.

No campo dos sistemas de recomendação para corrida, os estudos elencados evidenciam que

sistemas de recomendação têm grande potencial de apoio à jornada do atleta recreacional, mas também revelam lacunas: grande parte das abordagens concentra-se em recomendações baseadas em similaridade ou inferência de condicionamento físico, e não em otimização explícita de variáveis controláveis. O presente estudo contribui para preencher esse vazio ao empregar algoritmos de otimização para transformar previsões de desempenho em recomendações diretamente acionáveis, oferecendo ao corredor sugestões fundamentadas em mudanças possíveis em seus padrões de treino. Assim, enquanto Smyth et al. [2022] focam em casos históricos e similaridade, o modelo proposto direciona-se à identificação matemática de ajustes ótimos, considerando simultaneamente restrições e características próprias de cada *cluster*. Trata-se, portanto, de uma abordagem orientada à ação, cujo objetivo é indicar caminhos factíveis de aprimoramento, em contraste com recomendações derivadas apenas de padrões comparativos.

Em síntese, o trabalho aqui desenvolvido posiciona-se como uma contribuição integrada e inovadora na literatura ao combinar clusterização, modelagem preditiva e otimização, propondo um fluxo analítico capaz de capturar perfis de corredores, estimar desempenho com base em múltiplas fontes de dados e gerar recomendações personalizadas suportadas por algoritmos matemáticos. Essa integração, ainda pouco explorada, aponta para novas possibilidades de sistemas inteligentes de apoio ao treinamento esportivo, especialmente para corredores recreacionais que buscam evolução estruturada com base em seus próprios dados. A Tabela III.2 consolida as principais comparações do estudo proposto com outros estudos, elencando suas principais contribuições.

Tabela III.2: Contribuições do presente estudo em relação à literatura

Temática	**Literatura**	**Presente Estudo**
Análise de desempenho	Concentram-se em abordagens isoladas, seja apenas na caracterização de perfis, na previsão de performance, ou na proposição de sistemas de recomendação. Além disso, são raros os estudos que integram informações fornecidas por meio de pesquisa com histórico de treinamento captado por dispositivos vestíveis.	Integração sistemática que envolve clusterização, previsão de desempenho e recomendação, utilizando dados que consideram tanto as características e informações fornecidas por meio de pesquisa, quanto histórico de treinamento captado por dispositivo vestível.

Tabela III.2 (continuação)

Temática	**Literatura**	**Presente Estudo**
Clusterização na corrida	Algoritmos de clusterização aplicados para descrever perfis.	Algoritmos de clusterização aplicados tanto para a descrição de perfis quanto para a segmentação da base de dados, a qual serve como *input* para modelos preditivos específicos a cada perfil.
Previsão de desempenho na corrida	Focados em populações específicas, restringindo a generalização dos achados.	Capaz de prever o desempenho tanto de forma geral, quanto segmentada por grupos específicos de corredores.
Recomendações no esporte	Concentram-se em recomendações baseadas em similaridade ou através de inferência de condicionamento físico.	Propõe a recomendação através da otimização explícita de variáveis que podem ser controladas, considerando simultaneamente restrições e características próprias de cada *cluster*.

Capítulo IV Metodologia

Este capítulo apresenta uma visão geral das principais etapas envolvidas no desenvolvimento da análise, abrangendo desde a estruturação inicial do conjunto de dados até a aplicação dos modelos preditivos e do processo de recomendação. Descreve-se o procedimento de coleta das informações, bem como as etapas de enriquecimento e limpeza necessárias para garantir consistência, completude e qualidade analítica. Adicionalmente, explicita-se o processo de seleção das variáveis utilizadas nas modelagens, fundamentado em estudos prévios e em referências consolidadas na literatura, assegurando que os preditores escolhidos possuíssem respaldo teórico e relevância prática. Também são expostas as metodologias empregadas para a construção dos modelos de clusterização, previsão de desempenho e formulação das recomendações baseadas em variáveis controláveis. Por fim, são apresentadas as abordagens utilizadas para interpretar os dados processados e validar os resultados, assegurando que as análises produzidas sejam robustas, coerentes e alinhadas aos objetivos do estudo.

A Seção IV.1 detalha as fontes utilizadas, os critérios de extração das informações e os procedimentos adotados para a consolidação do conjunto de dados empregado no estudo. Na sequência, a Seção IV.2 descreve as etapas de enriquecimento e limpeza, incluindo tratamentos de inconsistências, padronizações e integrações necessárias para preparar a base analítica. Por fim, a Seção IV.3 apresenta a aplicação dos algoritmos de clusterização, previsão de desempenho e otimização, explicitando as configurações adotadas, o fluxo de execução e a forma como essas técnicas foram integradas para sustentar a geração das recomendações propostas.

IV.1 Coleta dos Dados

Foi conduzida uma pesquisa com corredores de diferentes níveis de maturidade esportiva, estruturada a partir de um questionário que reuniu perguntas sobre informações pessoais, ocupacionais, e de hábitos de treino. Complementarmente, os participantes puderam, ao final do questionário, compartilhar seus registros históricos de atividades provenientes do aplicativo Strava[1], o qual fornecem dados detalhados em séries temporais sobre as corridas realizadas, incluindo métricas como ritmo, localidade, distância e elevação. Todos esses dados foram utilizados para compor a base

[1]Disponível em: `https://www.strava.com/`. Último acesso em: 25 de abril de 2025.

empírica do estudo.

O instrumento de coleta consistiu em um formulário eletrônico elaborado no *Google Forms*, enviado individualmente a cada participante, com o objetivo de reunir informações sobre características pessoais, histórico de prática esportiva e hábitos relacionados ao treinamento de corrida. Não foram estabelecidos critérios de idade ou de experiência prévia em corrida como requisito de participação.

O questionário foi dividido em três seções principais. A primeira contém oito perguntas de caráter pessoal, abrangendo informações demográficas e físicas, como nome, sexo, data de nascimento, peso, altura e situação ocupacional. Também foram incluídas questões sobre condições de saúde ou restrições físicas que pudessem limitar a prática da corrida, bem como o histórico de lesões relacionadas à atividade.

A segunda seção aborda as práticas de corrida dos participantes, por meio de oito questões voltadas a aspectos comportamentais e técnicos. Entre os tópicos explorados estavam o tempo de prática (quando o atleta começou a correr), a frequência semanal de treinos, o tipo de corrida preferido, o acompanhamento por profissionais, o uso de aplicativos para registro de treinos, a experiência prévia em provas de corrida e as preferências quanto ao tipo de companhia (sozinho ou acompanhado) e ambiente de treino.

Por fim, a terceira seção trata de informações adicionais sobre a rotina esportiva dos participantes. As quatro perguntas dessa seção buscaram compreender se o atleta já havia competido em outras modalidades, se praticava outros tipos de exercício físico, com que frequência o fazia e se esses treinos também eram registrados em aplicativos.

Além das respostas ao questionário, os participantes tiveram a opção de compartilhar seu histórico de treino, contendo dados detalhados sobre suas atividades registradas no Strava. Esse material foi essencial para o enriquecimento da base de dados, permitindo análises mais aprofundadas sobre padrões de treino e desempenho.

Todos os participantes consentiram formalmente com a utilização de seus dados por meio da assinatura do Termo de Consentimento Livre e Esclarecido (TCLE), em conformidade com as diretrizes do Comitê de Ética em Pesquisa. O estudo foi aprovado pelo Comitê de Ética em Pesquisa do Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro (IFRJ), sob o Certificado de Apresentação para Apreciação Ética (CAAE) nº 85859024.5.0000.5268. Todas as etapas da investigação foram conduzidas em conformidade com a legislação vigente e com os requisitos institucionais aplicáveis, assegurando a integridade científica e o respeito aos princípios éticos. Previamente à participação, todos os voluntários receberam informações claras sobre os objetivos e procedimentos da pesquisa, e forneceram formalmente o consentimento para a utilização de seus dados exclusivamente para fins acadêmicos e científicos.

IV.2 Pré-processamento dos Dados

Os dados coletados foram processados na linguagem de programação Python (versão 3.13.1), escolhida por sua ampla utilização em ciência de dados e aprendizado de máquina. Para leitura e manipulação das informações provenientes dos registros de atividades, foram utilizadas bibliotecas específicas para cada formato de arquivo: Gpxo[2] para arquivos no formato GPX, Fitdecode[3] para arquivos FIT e Tcxreader[4] para arquivos TCX. Essas bibliotecas permitiram a extração estruturada dos dados brutos de treino, viabilizando as etapas subsequentes de tratamento e análise. Já as respostas do questionário foram lidas e manipuladas utilizando bibliotecas consolidadas em Python, como Pandas[5], que possibilitam a organização, limpeza e exploração eficiente de dados tabulares. A Figura IV.1 ilustra, de maneira simples, como esse fluxo de obtenção dos dados ocorreu.

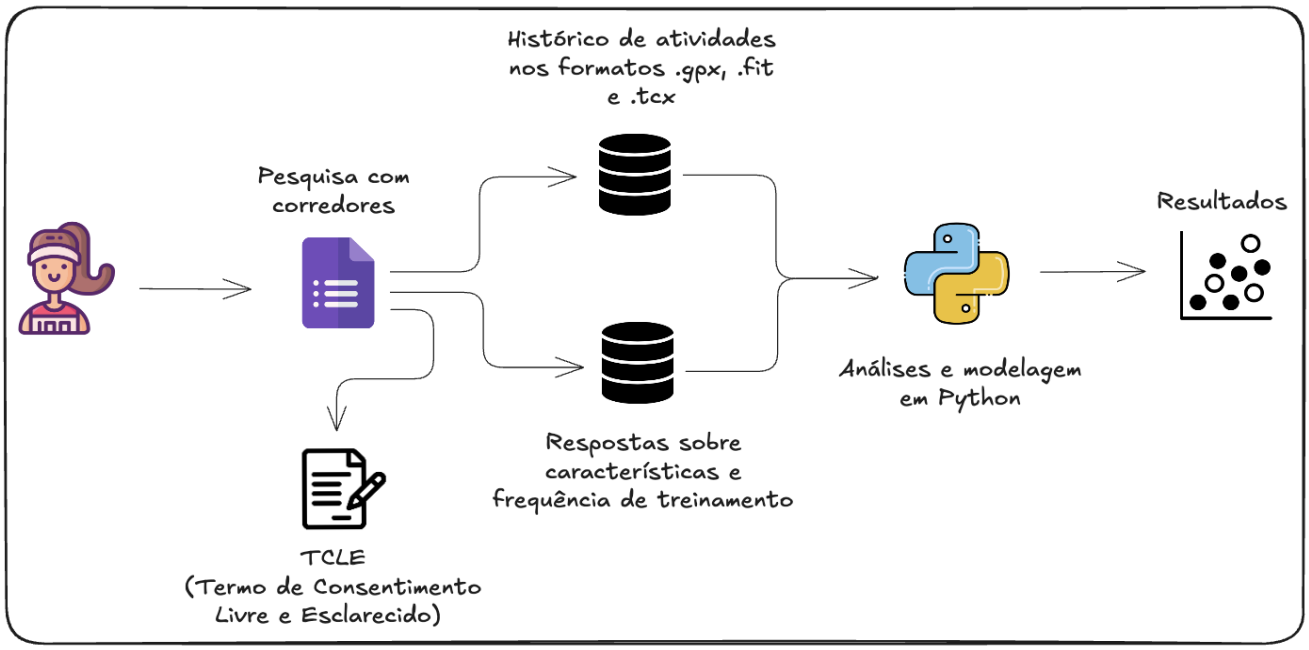


Figura IV.1: Fluxo resumido para obtenção dos dados

IV.2.1 Geração e Enriquecimento de Variáveis

A partir das respostas obtidas por meio do questionário, foram extraídas, ao todo, 21 variáveis originais, abrangendo aspectos pessoais, de prática de corrida e de rotina de treinamento, que foram declaradas pelos respondentes. Adicionalmente, a partir do histórico de treinamento disponibilizado pelos atletas, foram extraídas 8 variáveis destinadas à caracterização dos treinos. Cabe destacar que os registros de treinamento são, em sua maioria, fornecidos em intervalos de um segundo, de

[2]Disponível em: `https://pypi.org/project/gpxo/`. Último acesso em: 17 de maio de 2025.
[3]Disponível em: `https://fitdecode.readthedocs.io/en/latest/`. Último acesso em: 17 de maio de 2025.
[4]Disponível em: `https://pypi.org/project/tcxreader/`. Último acesso em: 17 de maio de 2025.
[5]Disponível em: `https://pandas.pydata.org/`. Último acesso em: 08 de outubro de 2025.

modo que cada treino, para cada atleta, é composto por múltiplas observações, ou seja, cada linha da base de dados representa um instante específico do treino.

Na Tabela IV.1, apresentam-se todas as variáveis **originais** utilizadas no estudo, acompanhadas de suas respectivas fontes e tipos de dados.

Tabela IV.1: Variáveis originais extraídas do questionário e do histórico de atividades

Nº	**Variável**	**Fonte**	**Tipo de variável**	**Nível da variável**
1	Data do preenchimento	Questionário	Data	Atleta
2	Nome	Questionário	Categórica	Atleta
3	Sexo	Questionário	Categórica	Atleta
4	Data de nascimento	Questionário	Data	Atleta
5	Peso	Questionário	Numérica	Atleta
6	Altura	Questionário	Numérica	Atleta
7	Ocupação	Questionário	Categórica	Atleta
8	Problemas de saúde	Questionário	Categórica	Atleta
9	Data de lesão	Questionário	Data	Atleta
10	Data aproximada de início na corrida	Questionário	Data	Atleta
11	Frequência de treinos de corrida na semana	Questionário	Numérica	Atleta
12	Tipo favorito de corrida	Questionário	Categórica	Atleta
13	Auxílio profissional	Questionário	Categórica	Atleta
14	Uso de dispositivos eletrônicos (app)	Questionário	Categórica	Atleta
15	Se já fez alguma prova de corrida	Questionário	Categórica	Atleta
16	Se prefere correr sozinho ou acompanhado	Questionário	Categórica	Atleta
17	Tipo favorito de terreno para corrida	Questionário	Categórica	Atleta
18	Prática de outros exercícios	Questionário	Categórica	Atleta

Tabela IV.1 (continuação)

Nº	Variável	Fonte	Tipo de variável	Nível da variável
19	Frequência de prática de outros exercícios na semana	Questionário	Numérica	Atleta
20	Se já competiu em outro esporte	Questionário	Categórica	Atleta
21	Qual outro esporte já competiu	Questionário	Categórica	Atleta
22	*Timestamp*	Histórico de atividades	Data/Hora	Atleta/Treino
23	Latitude	Histórico de atividades	Numérica	Atleta/Treino
24	Longitude	Histórico de atividades	Numérica	Atleta/Treino
25	Distância (km)	Histórico de atividades	Numérica	Atleta/Treino
26	Velocidade (km/h)	Histórico de atividades	Numérica	Atleta/Treino
27	Cadência	Histórico de atividades	Numérica	Atleta/Treino
28	Elevação	Histórico de atividades	Numérica	Atleta/Treino
29	Bússola	Histórico de atividades	Numérica	Atleta/Treino

Em um segundo momento, procedeu-se o enriquecimento do conjunto de dados com o objetivo de ampliar as possibilidades analíticas e otimizar a interpretação dos resultados. Foram desenvolvidas as seguintes variáveis derivadas:

- **Idade do atleta**: variável calculada a partir da diferença entre a data de nascimento e a data de preenchimento do questionário, permitindo a realização de análises por faixa etária e a investigação de possíveis correlações com outras variáveis de interesse.

- **Tempo total desde o início das atividades de corrida**: corresponde ao número de meses decorridos desde a data declarada de início nas atividades de corrida, possibilitando mensurar a experiência do atleta na modalidade.
- **Tempo total desde a data de lesão**: representa o número de meses transcorridos desde a data declarada da última lesão. Caso o atleta não tenha relatado histórico de lesão, o valor atribuído é igual a zero.
- **Histórico de lesão**: variável booleana, assumindo valor 1 para atletas que relataram alguma lesão e 0 para aqueles que não apresentaram histórico.
- **Prática de musculação**: variável booleana derivada da informação sobre outros exercícios praticados, assumindo valor 1 caso o atleta pratique musculação regularmente e 0 caso contrário.
- **Auxílio profissional**: variável transformada de formato categórico para numérico, onde as respostas "Sim", "Não" e "Não, mas já corri" foram convertidas em 1, 0 e 0,5, respectivamente, com o intuito de mensurar gradualmente o nível de suporte técnico recebido pelo atleta.
- **Data do treino**: variável extraída do campo *timestamp* do histórico de atividades, utilizada para a integração entre diferentes bases de dados (*joins*) e para a organização temporal das sessões de treino.
- **Idade do atleta no treino**: variável calculada pela diferença entre a data de nascimento e a data do treino, permitindo análises longitudinais da evolução do desempenho em diferentes faixas etárias.
- **Total de treinos do atleta**: representa o número total de treinos registrados por atleta, possibilitando avaliar a consistência e frequência de prática.
- **Velocidade máxima do treino (km/h)**: indica a maior velocidade atingida pelo atleta em cada sessão de treino.
- **Velocidade mínima do treino (km/h)**: indica a menor velocidade registrada durante o treino, podendo refletir momentos de pausa, aquecimento ou recuperação.
- **Desvio-padrão da velocidade do treino**: métrica que quantifica a variabilidade da velocidade ao longo do treino, fornecendo indícios sobre a regularidade do ritmo de corrida.
- **Número do treino**: sequência numérica que identifica a ordem cronológica de cada treino realizado por atleta.

- **Número do treino dentro do mês**: contagem sequencial de treinos realizados dentro de um mesmo mês, útil para análises de periodicidade e carga mensal.
- **Número do treino dentro da semana**: contagem sequencial de treinos realizados dentro da semana, permitindo avaliar a frequência semanal de prática.
- **Dias desde o último treino**: variável calculada pela diferença em dias entre a data do treino atual e a data do treino anterior, refletindo o intervalo de recuperação entre sessões.
- **Condição de lesão no treino**: variável booleana que indica se o atleta encontrava-se em período de lesão no momento do treino. Para essa classificação, considerou-se o mês da lesão declarada e o mês subsequente como período ativo de lesão.
- **Tipo de treino**: variável derivada a partir do desvio-padrão da velocidade, utilizada como uma *proxy* para identificar treinos intervalados. Adotou-se a seguinte regra de classificação: valores superiores a 3 indicam treinos intervalados (1), enquanto valores iguais ou inferiores indicam treinos contínuos (0).
- **Velocidade média do treino anterior (km/h)**: utilizando o conceito de janela deslizante (*rolling window*), essa variável representa a velocidade média registrada no treino imediatamente anterior ($t-1$).
- **Velocidade média do treino antes do anterior (km/h)**: análoga à variável anterior, porém referente ao treino realizado dois ciclos antes ($t-2$).
- **Velocidade média do treino antes da antes do anterior (km/h)**: variável calculada com base no treino realizado três ciclos antes ($t-3$), permitindo capturar tendências temporais de desempenho.
- **Distância do treino anterior (km)**: variável também derivada do conceito de janela deslizante, representando a distância total percorrida no treino imediatamente anterior ($t-1$).

Além dessas variáveis declaradas e calculadas, foram também utilizados dados de temperatura obtidos por meio da biblioteca Meteostat[6], que fornece séries históricas de variáveis meteorológicas com cobertura global. Essa informação pode ser obtida a partir das coordenadas geográficas (latitude e longitude), permitindo associar as condições climáticas específicas aos locais onde os treinos foram realizados.

Para cada treino, foram identificados os pontos inicial e final de latitude e longitude registrados no histórico da atividade, e, a partir dessas coordenadas, foram obtidos os valores correspondentes

[6]Disponível em: `https://pypi.org/project/meteostat/`. Último acesso em: 21 de outubro de 2025.

de temperatura ambiente. Em seguida, calculou-se a média entre as temperaturas do início e do fim do percurso, de modo a estimar a temperatura média do treino. Essa variável passou, então, a compor a base de dados, uma vez que a temperatura ambiente é reconhecida como um fator determinante no desempenho em corrida, influenciando a termorregulação, o gasto energético e a percepção de esforço dos atletas. Variações térmicas podem afetar significativamente a velocidade, a resistência e o tempo de fadiga durante treinos e competições, tornando os dados de temperatura uma variável preditiva relevante para a análise de desempenho [El Helou et al., 2012; Ely et al., 2008].

Na Tabela IV.2, são apresentadas as variáveis **derivadas**, obtidas a partir do processamento e transformação das variáveis originais e as variáveis obtidas de fontes externas.

Tabela IV.2: Variáveis derivadas criadas para análise

Nº	**Variável**	**Tipo de variável**	**Nível da variável**
1	Idade do atleta	Numérica	Atleta
2	Tempo total desde o início das atividades de corrida	Numérica	Atleta
3	Tempo total desde a data de lesão	Numérica	Atleta
4	Histórico de lesão	Booleana	Atleta
5	Prática de musculação	Booleana	Atleta
6	Auxílio profissional	Numérica	Atleta
7	Data do treino	Data	Atleta/Treino
8	Idade do atleta no treino	Numérica	Atleta/Treino
9	Total de treinos do atleta	Numérica	Atleta
10	Velocidade máxima do treino (km/h)	Numérica	Atleta/Treino
11	Velocidade mínima do treino (km/h)	Numérica	Atleta/Treino
12	Desvio-padrão da velocidade do treino	Numérica	Atleta/Treino
13	Número do treino	Numérica	Atleta/Treino
14	Número do treino dentro do mês	Numérica	Atleta/Treino
15	Número do treino dentro da semana	Numérica	Atleta/Treino
16	Dias desde o último treino	Numérica	Atleta/Treino
17	Condição de lesão no treino	Booleana	Atleta/Treino
18	Tipo de treino	Booleana	Atleta/Treino
19	Velocidade média do treino anterior (km/h)	Numérica	Atleta/Treino
20	Velocidade média do treino antes do anterior (km/h)	Numérica	Atleta/Treino
21	Velocidade média do treino antes da antes do anterior (km/h)	Numérica	Atleta/Treino
22	Distância do treino anterior (km)	Numérica	Atleta/Treino
23	Temperatura no treino (°C)	Numérica	Atleta/Treino

Sendo assim, com base nos dados coletados e enriquecidos, foram consolidadas duas bases distintas: uma **base questionário**, estruturada no nível do atleta, contendo informações demográficas,

de perfil e histórico esportivo; e uma **base atividades**, organizada no nível treino/atleta, compreendendo registros detalhados das sessões de exercício. A vinculação entre essas bases é estabelecida mediante a chave comum de nome do atleta.

IV.2.2 Limpeza dos Dados

Apesar de minoria, nem todos os registros apresentaram valores devidamente preenchidos, especialmente na base atividades, o que exigiu a aplicação de procedimentos de tratamento e limpeza dos dados, a fim de garantir sua consistência e adequação para as análises subsequentes. Nesse sentido, foram adotados alguns critérios de exclusão de alguns treinos. Primeiramente, eliminaram-se treinos com velocidades muito elevadas (acima de 30 km/h), uma vez que não se mostravam compatíveis com a prática de corrida, possivelmente resultando de erros de registro ou atividades realizadas em outros meios de transporte. Também foram removidos treinos cujas características sugeriam práticas incompatíveis, como natação. Nesse caso, consideraram-se para exclusão atividades com distância total em torno de 1,5 km. Por fim, foram desconsiderados registros em que a velocidade permaneceu inteiramente nula ao longo de todo o treino, indicando inconsistências ou falhas na sincronização dos dados. Esses procedimentos de filtragem foram fundamentais para garantir a qualidade da base final de dados, reduzindo ruídos e prevenindo vieses nas análises subsequentes.

Após a aplicação dos critérios de limpeza mencionados, procedeu-se ao tratamento dos valores ausentes identificados nas variáveis remanescentes. Para os casos em que havia registros nulos, adotou-se o método de imputação pela média, de modo a preservar a completude da base sem comprometer a consistência estatística dos dados. Sempre que possível, esse preenchimento foi realizado respeitando o mesmo período temporal, a fim de manter a coerência contextual das medições e minimizar possíveis distorções decorrentes de variações sazonais ou temporais. No caso específico da variável de Temperatura no treino, mesmo após a aplicação dos procedimentos de imputação, algumas poucas observações permaneceram sem valor disponível. Para esses casos excepcionais, optou-se por atribuir o valor fixo de 20°C, considerado uma temperatura neutra e representativa de condições típicas de prática esportiva, garantindo assim a integridade e a comparabilidade das análises subsequentes.

Por fim, após a etapa de agregação das métricas por atleta e por sessão de treino, foi identificado um pequeno número de registros duplicados (17 treinos) que, embora apresentassem o mesmo identificador de arquivo de origem, estavam associados a datas distintas. Esse comportamento sugere a ocorrência de inconsistências pontuais no processo de registro ou exportação das atividades pelo aplicativo utilizado, possivelmente também decorrentes de erros de sincronização ou salvamento. Considerando o potencial impacto desses registros duplicados sobre a qualidade das análises, tais observações foram excluídas do conjunto final de dados.

IV.3 Modelagem: As 3 Fases

A etapa de modelagem foi conduzida em três fases principais, com o objetivo de compreender os diferentes perfis de corredores, prever o desempenho esportivo e, posteriormente, identificar os fatores que mais contribuem para sua maximização.

Inicialmente, foi aplicada uma análise de clusterização com o intuito de identificar grupos de atletas com características semelhantes. Para essa segmentação, não foram utilizadas informações da base atividades (histórico de treinamento), mas apenas algumas das variáveis da base questionário. Essa escolha permitiu observar padrões de comportamento e características individuais entre corredores com diferentes perfis físicos, hábitos autorrelatados e histórico esportivo, fornecendo uma visão exploratória sobre a heterogeneidade da amostra.

Em seguida, foi desenvolvido um modelo preditivo voltado à estimativa do desempenho dos atletas. Nessa fase, foi treinado um modelo geral e, adicionalmente, um modelo específico para cada *cluster* identificado na fase anterior. Os modelos consideraram tanto as variáveis da base questionário quanto as informações derivadas da base atividades, incluindo métricas de características, frequência, intensidade e consistência, com o objetivo de compreender como o conjunto de informações pessoais e de treinamento se relacionavam com o desempenho observado.

Por fim, foi implementado um algoritmo otimizador, cujo propósito foi determinar as combinações de variáveis controláveis que maximizavam a variável de interesse, isto é, o desempenho do atleta, também para cada *cluster* identificado. Essa abordagem permitiu explorar cenários simulados e identificar quais ajustes nos padrões de treino poderiam potencialmente levar a melhores resultados, contribuindo para recomendações mais individualizadas e embasadas.

A Figura IV.2 ilustra como o fluxo de treinamento dos modelos de aprendizado de máquina ocorreu.

IV.3.1 Clusterização de Corredores

Conforme mencionado anteriormente, o objetivo da etapa de clusterização consistiu em segmentar os usuários a partir de suas características individuais declaradas, sem considerar informações provenientes do histórico de treinamento. Essa abordagem buscou identificar grupos de corredores com perfis semelhantes em termos de atributos pessoais e experiência esportiva, possibilitando uma compreensão mais estruturada das diferenças entre perfis de atletas. Para tanto, foram utilizadas quatro variáveis extraídas da base questionário, a nível atleta: idade do atleta, peso, altura e tempo total desde o início das atividades de corrida. Essas variáveis foram selecionadas por refletirem aspectos fundamentais do condicionamento físico, da composição corporal e da trajetória esportiva dos participantes, elementos potencialmente relevantes para a caracterização dos grupos formados.

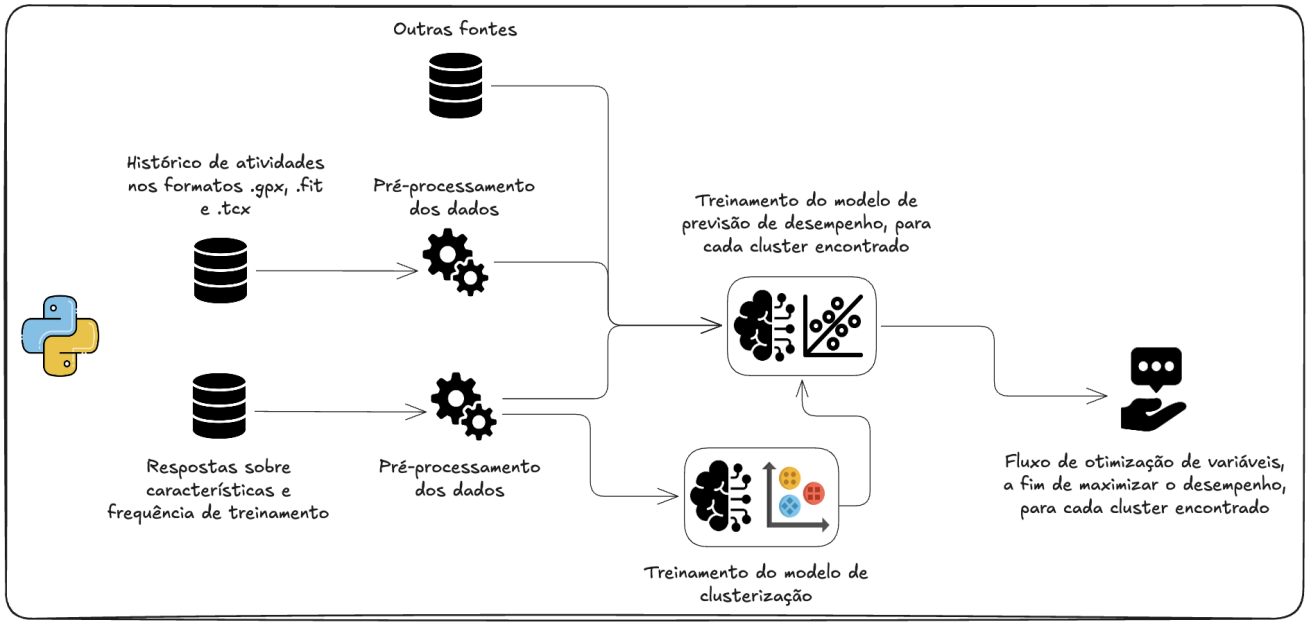


Figura IV.2: Fluxo de treinamento dos modelos

Antes da aplicação dos algoritmos de clusterização, procedeu-se à padronização das variáveis, com o intuito de eliminar a influência de diferenças de escala e assegurar que todas contribuíssem de maneira equitativa para a formação dos agrupamentos. Essa etapa foi conduzida por meio da classe `StandardScaler`[7], disponível na biblioteca *scikit-learn*[8], a qual transforma as variáveis para possuírem média zero e desvio padrão unitário. A padronização é uma prática essencial em métodos baseados em distância, uma vez que impede que variáveis de maior magnitude exerçam influência desproporcional sobre o processo de agrupamento.

Seguindo práticas amplamente adotadas na literatura para redução de dimensionalidade e mitigação de redundâncias entre variáveis em bases de dados multivariadas na etapa de agrupamento, foi aplicado o método de *Principal Component Analysis* (PCA)[9] nos dados já padronizados. O PCA transforma um conjunto de variáveis possivelmente correlacionadas em um novo conjunto de componentes independentes, que retêm a maior parte da variabilidade explicada pelos dados originais. No presente estudo, a redução foi conduzida de modo que aproximadamente 80% da variabilidade total dos dados fosse preservada, percentual comumente adotado na literatura para garantir a manutenção da estrutura informacional essencial sem comprometer a interpretabilidade dos resultados. Essa transformação permite eliminar redundâncias e destacar as dimensões mais relevantes para a diferenciação dos grupos, tornando processos como os de clusterização mais robustos e interpretáveis. Essa abordagem é consistente com trabalhos anteriores que investigaram a clusterização no

[7]Disponível em: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`. Último acesso em: 08 de novembro de 2025.
[8]Disponível em: `https://scikit-learn.org/stable/`. Último acesso em: 08 de novembro de 2025.
[9]Disponível em: `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html/`. Último acesso em: 01 de novembro de 2025.

contexto esportivo, os quais também incorporaram o uso do PCA em suas etapas de análise, como demonstrado em Jauhiainen et al. [2020], Koegel et al. [2024] e Rodriguez Rivadulla et al. [2024].

Na sequência, foram aplicados três algoritmos amplamente reconhecidos na literatura de aprendizado não supervisionado: *K-means*, *Agglomerative Clustering* e DBSCAN, todos implementados na biblioteca *scikit-learn*. A execução dos modelos foi realizada, respectivamente, por meio das classes `KMeans`, `AgglomerativeClustering` e `DBSCAN`. A adoção desses três métodos teve como propósito abarcar diferentes pressupostos de formação de grupos (particional, hierárquico e baseado em densidade) permitindo uma análise comparativa mais abrangente acerca das possíveis estruturas de segmentação presentes nos dados.

Para o algoritmo *K-means*, a determinação do número ótimo de agrupamentos foi conduzida por meio do método do cotovelo[10], o qual consiste na identificação do ponto de inflexão na curva que relaciona o número de *clusters* com a soma dos erros quadráticos intra-grupo (*Within-Cluster Sum of Squares* (WCSS)). Esse ponto representa o equilíbrio entre a complexidade do modelo e o ganho explicativo proporcionado pela divisão adicional dos grupos, evitando tanto o subajuste quanto o sobreajuste do modelo.

No caso do *Agglomerative Clustering*, adotou-se uma abordagem hierárquica aglomerativa com limiar de distância (*distance threshold*) igual a 5, sem a definição prévia do número de agrupamentos. Essa configuração permite que as instâncias sejam progressivamente unidas com base em suas distâncias de similaridade até que o limite estabelecido seja atingido, resultando em uma estrutura hierárquica representada sob a forma de dendrograma. Para o critério de ligação, foi utilizado o método de *Ward*, amplamente empregado na literatura por minimizar a variância intra-grupo a cada etapa de fusão, produzindo agrupamentos mais compactos e homogêneos. Tal representação possibilita identificar relações de proximidade entre os indivíduos e observar diferentes níveis de granularidade na formação dos grupos.

Por sua vez, o algoritmo DBSCAN foi explorado a partir de uma busca sistemática nos hiperparâmetros `eps` (raio de vizinhança) e `min_samples` (número mínimo de pontos necessários para formar um *cluster*). Foram testadas combinações de valores no intervalo de 0,1 a 2 para o parâmetro `eps` e de 3 a 10 para o parâmetro `min_samples`. Para cada combinação, calculou-se o índice de silhueta[11], sendo descartadas aquelas configurações que resultaram em apenas um agrupamento ou em conjuntos compostos exclusivamente por ruído. A configuração final do modelo DBSCAN foi selecionada com base no maior valor do coeficiente de silhueta obtido, garantindo a formação de agrupamentos coesos e bem separados.

Por fim, a métrica de silhueta foi empregada para a comparação quantitativa do desempenho dos

[10]Disponível em: `https://kneed.readthedocs.io/en/stable/parameters.html`. Último acesso em: 01 de novembro de 2025.

[11]Disponível em: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`. Último acesso em: 22 de outubro de 2025.

três algoritmos, por fornecer uma medida simultânea de compacidade intra-grupo e separação inter-grupos. Adicionalmente, foi avaliado o índice de Davies–Bouldin[12], que complementa essa análise ao quantificar a similaridade média entre os *clusters*, sendo desejáveis valores menores e, portanto, útil como critério adicional de validação. Após a seleção do algoritmo com melhor desempenho segundo essas métricas, foram conduzidos testes estatísticos complementares com o objetivo de verificar a significância das diferenças entre os *clusters* identificados, assegurando que as segmentações representassem distinções reais e estatisticamente relevantes entre os perfis dos corredores analisados.

IV.3.2 Previsão de Desempenho

Para a previsão de desempenho, foram avaliados dois modelos de regressão: a Regressão Linear Múltipla, implementada por meio da classe `LinearRegression` da biblioteca *scikit-learn*, e o modelo XGBoost da biblioteca *XGBoost*[13], um método de aprendizado de máquina baseado em *gradient boosting* com árvores de regressão, amplamente adotado na literatura recente. A inclusão desses modelos se justifica, também, pelo amplo uso de abordagens de regressão na literatura relacionada à análise e predição de desempenho esportivo, conforme sintetizado na Tabela III.1. Nesse contexto, a velocidade média foi adotada como uma *proxy* de desempenho, por refletir de forma agregada a eficiência e a intensidade do treino realizado, sendo frequentemente utilizada em estudos que buscam estimar ou comparar o rendimento de atletas em atividades de corrida.

A utilização de dois modelos com características distintas permitiu explorar diferentes formas de relação entre as variáveis explicativas e a variável resposta contínua. Enquanto a Regressão Linear possibilita a estimação da contribuição individual dos fatores e favorece a interpretação dos coeficientes e do ajuste global, o XGBoost apresenta maior flexibilidade para capturar relações não lineares e interações complexas entre as variáveis, contribuindo para a comparação entre abordagens com diferentes níveis de interpretabilidade e capacidade preditiva.

A base de dados utilizada para o treinamento dos modelos foi construída a partir da integração entre a base questionário e a base atividades, conforme descrito nas seções anteriores. O grão da tabela de análise corresponde a cada treino realizado por cada atleta, de modo que as observações representam unidades temporais e individuais distintas. As variáveis utilizadas foram calculadas em dois níveis: a nível de treino e a nível de atleta, cujas descrições detalhadas podem ser consultadas nas Tabelas IV.1 e IV.2, respectivamente.

Então, em resumo, a variável alvo (*target*) dos modelos é a velocidade média registrada em cada treino e as variáveis de entrada incluem características tanto do treino em questão, quanto do atleta, sendo elas: velocidade máxima do treino, velocidade mínima do treino, distância percorrida,

[12]Disponível em: `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html`. Último acesso em: 10 de janeiro de 2026.

[13]Disponível em: `https://xgboost.readthedocs.io/en/stable/index.html` Último acesso em: 04 de janeiro de 2026.

número do treino dentro do mês, número do treino dentro da semana, auxílio profissional, dias desde o último treino, velocidade média do treino anterior, velocidade média do treino antes do anterior, velocidade média do treino antes da antes do anterior, distância do treino anterior, idade do atleta no treino, peso, altura, frequência de prática de outros exercícios na semana, total de treinos do atleta, tempo total desde o início das atividades de corrida, desvio-padrão da velocidade do treino, temperatura no treino, prática de musculação, histórico de lesão e tipo do treino.

Entre as variáveis consideradas, destacam-se aquelas associadas à experiência do atleta, como o total de treinos realizados e o tempo total desde o início das atividades de corrida. A inclusão dessas variáveis fundamenta-se em evidências empíricas identificadas em estudo anterior, intitulado "Diferenciando Perfis de Corredores por Meio de Pontos de Mudança nos Treinos" [Tito et al., 2024]. Nesse trabalho, observou-se que corredores mais experientes apresentaram menor proporção de pontos de mudança em suas séries temporais de velocidade ($p < 0{,}05$), sugerindo maior controle rítmico e estabilidade no desempenho ao longo dos treinos. Esse achado sugere que a estabilidade observada em atletas experientes pode ser utilizada como referência na construção de indicadores de perfil, enquanto a maior variabilidade encontrada em corredores menos experientes pode sinalizar necessidade de ajustes mais frequentes e recomendações específicas para favorecer sua evolução.

Adicionalmente, esse resultado reforça que a experiência do atleta exerce influência direta sobre seu desempenho, uma vez que corredores mais experientes tendem a apresentar respostas mais previsíveis e eficientes aos estímulos de treino [Tito et al., 2024]. Esse aspecto é fundamental na modelagem preditiva de desempenho, pois indica que o nível de experiência deve ser considerado como uma variável explicativa, permitindo ajustar as estimativas de resultado de forma mais realista ao reconhecer que a consistência e a capacidade adaptativa do atleta podem evoluir com o tempo de prática e com a maturidade esportiva adquirida.

No entanto, além de desconsiderar os níveis de velocidade, esse trabalho não avançou na análise de progressos no desempenho ao longo do tempo. Para preencher essa lacuna, um estudo subsequente buscou superar essa limitação ao identificar corredores que apresentaram melhora de desempenho, utilizando a detecção de desvio de conceito em séries temporais de treino. Foram comparadas métricas de treinamento entre períodos que antecederam melhora e piora, verificando-se que, nos casos de evolução, os indicadores foram significativamente superiores, evidenciando padrões de treino associados ao progresso [Tito et al., 2025].

Os resultados desse segundo estudo, intitulado *"Drift Detection for Identifying Training Patterns Prior to Performance Improvement in Runners"* [Tito et al., 2025], sustenta a utilização de variáveis relacionadas às características de treinamento, como distância percorrida, número do treino dentro do mês, número do treino dentro da semana e o desvio-padrão da velocidade. O trabalho analisou séries históricas de desempenho de corredores para identificar padrões precedentes a momentos de

melhoria ou declínio, e concluiu que atletas que apresentaram melhora de desempenho possuíam valores significativamente maiores para indicadores de treinamento, incluindo maior distância média percorrida, maior frequência semanal e mensal de treinos, além de maior variação na velocidade média, todos com diferenças estatisticamente significativas ($p < 0{,}05$). Os resultados dos testes aplicados para a comparação entre os recortes encontram-se na Tabela IV.3, que também está exposta no estudo. A Figura IV.3, também presente no estudo, complementa essa análise ao ilustrar graficamente os valores mais elevados das métricas no recorte associado à melhora.

Tabela IV.3: Resultados dos testes dos indicadores de treino

Training indicators (avg)	Improved runners	Declining runners	p-value
Distance (km)	8.46	5.29	0.00***
Variation in training speed (km/h)	2.76	2.18	0.01*
Training sessions per week (freq.)	2.87	1.95	0.00***
Training sessions per month (freq.)	11.59	4.90	0.00*

***p < 0.0001; **p < 0.001; *p < 0.01

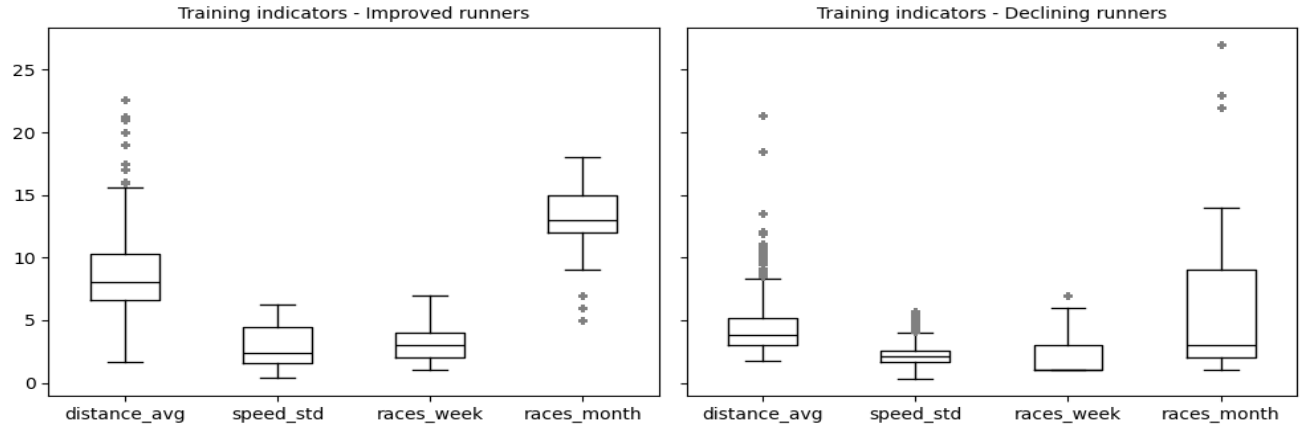


Figura IV.3: Distribuição dos indicadores de treinamento

Esses resultados sugerem que tanto o maior volume quanto a maior variabilidade dos treinos estão associados a progressões no desempenho, reforçando a importância dessas métricas como indicadores indiretos do engajamento e da carga de treino. As variáveis de treinamento que apresentaram relação significativa com a performance despontam, portanto, como potenciais preditores de sucesso, alinhando-se às evidências já documentadas na literatura [Knechtle et al., 2011; Tanda, 2011; Suwankan et al., 2024]. Dessa forma, a inclusão dessas variáveis no presente modelo busca capturar padrões de treinamento associados à evolução do desempenho, em consonância com os achados empíricos previamente identificados e com a literatura que relaciona volume, intensidade e consistência de treino à melhoria de performance em corredores.

A seleção das demais variáveis também foi fundamentada em estudos prévios e na literatura especializada, que apontam a relevância de fatores contextuais e históricos para a explicação do desempenho em atividades de corrida. Diversos trabalhos destacam a importância de variáveis

como condições climáticas, características antropométricas e indicadores de carga e recuperação, que influenciam diretamente o rendimento dos atletas e a variabilidade do desempenho ao longo do tempo, conforme sintetizado na Tabela III.1.

Para fins de modelagem, o conjunto de dados foi dividido em 80% para treino e 20% para teste, de modo a possibilitar a avaliação da capacidade preditiva dos modelos em dados não utilizados durante o ajuste. Assim como na fase de clusterização, antes do treinamento dos modelos de previsão, todas as variáveis contínuas foram padronizadas utilizando o `StandardScaler` da biblioteca *scikit-learn*. A padronização também é uma etapa essencial em modelos de regressão, pois garante que todas as variáveis estejam em uma mesma escala, evitando que diferenças de magnitude influenciem de forma desproporcional a estimação dos coeficientes. Além disso, esse procedimento contribui para a estabilidade numérica do modelo e facilita a interpretação comparativa entre os parâmetros estimados.

Os modelos foram inicialmente treinados de forma geral, considerando o conjunto completo de treinos e atletas. No entanto, reconhecendo que os indivíduos podem apresentar padrões distintos de comportamento e desempenho, também foram estimados modelos específicos para cada agrupamento de atletas encontrado, conforme identificado na fase de clusterização descrita anteriormente. Essa abordagem busca capturar melhor as características particulares de cada grupo, permitindo que o modelo aprenda relações mais representativas entre as variáveis explicativas e a velocidade média dentro de contextos fisiológicos e comportamentais semelhantes. Dessa forma, o processo de modelagem torna-se mais sensível às diferenças individuais, o que tende a aprimorar a capacidade preditiva e interpretativa do modelo em relação ao desempenho esportivo.

Após o ajuste dos modelos, foram calculadas métricas de desempenho com o intuito de avaliar e comparar sua qualidade preditiva, utilizando-se critérios comuns a ambas as abordagens. Dentre essas métricas, destaca-se o coeficiente de determinação R^2, que expressa a proporção da variabilidade da velocidade média explicada pelas variáveis independentes incluídas no modelo, e o *Root Mean Squared Error* (RMSE), que mede o desvio médio entre os valores observados e os valores preditos, indicando o quão próximos estão as previsões dos valores reais. Essas métricas permitem a comparação direta entre os modelos considerados, de modo que valores mais elevados de R^2 e menores de RMSE indicam melhor desempenho preditivo.

Adicionalmente, com o objetivo de obter uma avaliação mais robusta e menos dependente de uma única partição dos dados, foi aplicada a técnica de validação cruzada, que consiste em dividir o conjunto de dados em múltiplos subconjuntos, utilizando parte deles para treinamento e parte para validação de forma alternada, permitindo uma estimativa mais estável do desempenho do modelo. Considerando que o conjunto de dados contém múltiplas observações associadas a um mesmo indivíduo em diferentes pontos no tempo, adotou-se a validação cruzada do tipo *GroupKFold*,

por meio das rotinas `GroupKFold` e `cross_val_score` da biblioteca *scikit-learn*. Nessa abordagem, as partições são definidas com base em grupos, de modo que todas as observações de um mesmo indivíduo são alocadas integralmente em um único subconjunto, evitando assim o vazamento de informação entre os conjuntos de treinamento e validação.

No modelo geral, utilizou-se $k = 5$ partições, garantindo que, a cada iteração, um subconjunto distinto de indivíduos fosse empregado como conjunto de validação, enquanto os demais compunham o conjunto de treinamento. Já nos modelos estimados por *cluster* de corredores, o número de partições foi ajustado dinamicamente de acordo com o tamanho de cada *cluster*, de forma que o número de *splits* fosse igual ao número de indivíduos presentes no respectivo *cluster*, assegurando a viabilidade da validação cruzada mesmo em *clusters* com poucos elementos.

Como métrica de avaliação, utilizou-se o RMSE, calculado em cada uma das partições, sendo o desempenho final do modelo sintetizado pelo valor médio do RMSE obtido ao longo das iterações da validação cruzada. Essa estratégia proporciona uma estimativa mais estável da capacidade de generalização dos modelos, respeitando a estrutura longitudinal dos dados.

Além das métricas tradicionais de avaliação, foram também calculados os valores de explicabilidade baseados em SHAP *values*, com o objetivo de compreender de forma mais detalhada a contribuição individual de cada variável para as previsões dos modelos. Essa abordagem permite quantificar, para cada observação, quanto cada variável aumenta ou diminui a predição em relação ao valor médio, possibilitando uma análise interpretável mesmo em modelos com múltiplos preditores. Os SHAP *values* baseiam-se na teoria dos jogos cooperativos de Shapley (1953), atribuindo a cada variável uma importância proporcional à sua contribuição marginal para todas as combinações possíveis de preditores [Lundberg and Lee, 2017]. Essa técnica está disponível para uso em Python por meio da biblioteca SHAP[14], que oferece ferramentas gráficas e analíticas para visualização das contribuições das variáveis. Assim, além de fornecer uma medida global de relevância das variáveis, essa técnica permite visualizar e interpretar de forma intuitiva os fatores que mais influenciam a variável resposta, que pra este caso é a velocidade média, auxiliando na geração de *insights* sobre o comportamento e os determinantes do desempenho dos atletas.

Por fim, foi feita uma análise de sensibilidade com o objetivo de avaliar a influência marginal das variáveis explicativas sobre as predições do modelo de regressão escolhido. A análise foi conduzida por meio do método *one-at-a-time* (OAT), considerando exclusivamente as dez variáveis mais relevantes segundo a importância estimada pelos valores de SHAP. Para cada variável selecionada, foi aplicado individualmente um acréscimo equivalente a um desvio-padrão em relação aos seus valores originais, mantendo-se as demais variáveis constantes. A sensibilidade foi quantificada pela variação média na predição do modelo, o que permitiu identificar quais variáveis exercem maior impacto

[14]Disponível em: `https://shap.readthedocs.io/en/latest/`. Último acesso em: 08 de novembro de 2025.

marginal sobre a resposta prevista e complementando a interpretação dos resultados fornecida pelas métricas de importância do modelo.

IV.3.3 Recomendação para Maximização do Desempenho

Na fase de recomendação, foi empregada uma abordagem de otimização numérica com o objetivo de encontrar valores ótimos para variáveis controláveis que maximizassem a variável alvo do modelo de regressão escolhido, isto é, a velocidade média prevista. Para tal, utilizou-se a função `minimize` da biblioteca SciPy, que permite resolver problemas de otimização contínuos e restritos. O procedimento consistiu em ajustar as variáveis de entrada de modo a obter a maior predição possível para a velocidade média, conforme estimada pelo modelo treinado.

As variáveis controláveis consideradas no processo de otimização foram: velocidade máxima do treino, velocidade mínima do treino, distância percorrida, número do treino dentro da semana, frequência de prática de outros exercícios na semana, prática de musculação e tipo de treino. Essas variáveis foram selecionadas por representarem aspectos diretamente manipuláveis no contexto do treinamento esportivo, o que as torna adequadas para o processo de recomendação personalizada. Além disso, a escolha dessas variáveis foi corroborada pela análise de importância baseada nos valores SHAP, que indicou sua relevância na predição da variável resposta, reforçando seu papel explicativo no desempenho estimado pelos modelos. Assim, a otimização busca indicar ajustes realistas e factíveis nessas dimensões, respeitando as características individuais e os limites observados empiricamente nos dados.

A função objetivo foi definida para maximizar a velocidade média prevista pelo modelo de regressão, a partir da modificação dos valores dessas variáveis controláveis. Como a função `minimize` realiza a minimização de um valor escalar, o valor previsto foi multiplicado por -1, de forma que o mínimo da função correspondesse ao máximo da velocidade média. Essa estratégia é amplamente utilizada em procedimentos de otimização contínua, sendo computacionalmente eficiente e compatível com os algoritmos implementados na biblioteca SciPy.

Para garantir a plausibilidade e a coerência das recomendações, foram incorporadas restrições (*constraints*) ao processo de otimização, que asseguram o cumprimento de condições lógicas e limites físicos do problema. As restrições foram implementadas no formato de funções matemáticas, de modo que seus valores fossem não negativos para soluções viáveis. Foram consideradas três restrições principais: (i) a velocidade média prevista não poderia ultrapassar o valor máximo de velocidade registrado nos dados; (ii) a velocidade máxima otimizada deveria ser maior ou igual à velocidade média prevista, garantindo coerência entre as grandezas; e (iii) a distância percorrida otimizada deveria ser, no mínimo, igual à média observada, de forma a evitar recomendações incompatíveis com a realidade dos treinos analisados. Para lidar adequadamente com restrições de desigualdade

não lineares, utilizou-se o método SLSQP, disponível na implementação do SciPy, que combina eficiência computacional e robustez para esse tipo de problema.

Em complemento, foram definidos limites inferiores e superiores (*bounds*) para cada variável controlável, a fim de restringir o espaço de busca do algoritmo a valores factíveis e evitar extrapolações indesejadas. Esses limites foram determinados com base na média e no desvio-padrão observados para cada variável, adotando um intervalo de confiança de 95% (média $\pm$ 1,96 $\times$ desvio-padrão). Essa escolha permite que o algoritmo explore combinações plausíveis dentro do espectro empírico observado, mantendo o equilíbrio entre flexibilidade e realismo.

O ponto de partida da otimização (*initial point*) foi definido como a mediana observada de cada variável controlável, ajustada para garantir que o valor inicial respeitasse os limites estabelecidos. Essa configuração inicial oferece ao algoritmo uma posição de partida representativa e estável, reduzindo o risco de convergência para soluções não ótimas ou inviáveis.

Assim como o modelo de previsão, o processo de otimização foi aplicado em um primeiro momento usando o modelo generalista, e em seguida, para cada modelo treinado por perfil de atleta, conforme os agrupamentos identificados na etapa de clusterização. Essa decisão metodológica visa capturar as particularidades e padrões de comportamento específicos de cada grupo, possibilitando que as recomendações sejam mais personalizadas e aderentes ao perfil de desempenho dos corredores. Dessa forma, o procedimento de otimização não apenas busca maximizar o desempenho previsto, mas também respeita as diferenças individuais entre os atletas, garantindo maior precisão e aplicabilidade prática das recomendações geradas.

Para avaliação, também foi conduzida uma análise de sensibilidade local associada à etapa de otimização, com o objetivo de avaliar a robustez da solução ótima obtida e identificar quais variáveis de decisão exerceram maior influência marginal sobre a função objetivo. De forma análoga à análise de sensibilidade realizada na etapa de previsão, foi adotado o método OAT, no qual cada variável otimizada foi individualmente perturbada por um acréscimo equivalente a um desvio-padrão em torno do valor ótimo encontrado, mantendo-se as demais variáveis constantes. A sensibilidade foi mensurada pela variação na velocidade prevista pelo modelo, permitindo avaliar o impacto de pequenas variações nas decisões recomendadas. Essa análise é fundamental para verificar a estabilidade da solução ótima, identificar variáveis críticas que requerem maior controle e assegurar que as recomendações geradas pelo processo de otimização sejam robustas e aplicáveis em cenários reais, nos quais flutuações naturais nas variáveis são esperadas.

Por fim, e com o objetivo de avaliar os ganhos efetivamente atribuíveis à etapa de otimização, as previsões feitas pelos modelos foram comparadas com um cenário de referência (*baseline*). Esse *baseline* foi definido a partir da predição do próprio modelo, calculada sobre um perfil médio de atleta, no qual todas as variáveis explicativas foram fixadas em seus valores médios observados no

conjunto de dados, incluindo aquelas sujeitas à otimização. Dessa forma, as previsões do *baseline* refletem diretamente a resposta estimada pelo modelo para uma configuração média das variáveis, representando a ausência de qualquer intervenção ou ajuste direcionado no espaço de decisão. A comparação entre o valor ótimo e esse cenário de referência permite isolar o efeito da otimização sobre a resposta prevista, assegurando uma avaliação mais consistente e interpretável dos benefícios proporcionados pelo processo de otimização.

Capítulo V Resultados

Este capítulo apresenta os principais achados obtidos ao longo da investigação, reunindo análises quantitativas, interpretações substanciais e *insights* decorrentes da aplicação do modelo proposto. Além da avaliação da clusterização, da modelagem preditiva e do processo de otimização, inclui-se também a análise dos resultados da pesquisa realizada com os corredores, de modo a integrar percepções autorrelatadas, características individuais e hábitos de treino às evidências extraídas dos dados. São discutidos os contrastes entre diferentes perfis de atletas, os padrões identificados e as implicações práticas das recomendações geradas, oferecendo uma visão abrangente da eficácia e das contribuições do método empregado.

A Seção V.1 apresenta a análise descritiva e exploratória das respostas ao questionário, bem como das bases de treino fornecidas pelos participantes, evidenciando padrões comportamentais, características dos corredores e relações iniciais entre variáveis de interesse. Em seguida, a Seção V.2 reúne as evidências obtidas a partir das etapas analíticas do estudo: a subseção de clusterização discute a segmentação dos perfis de atletas e sua interpretabilidade; a subseção de previsão de desempenho avalia o poder preditivo dos modelos e a relevância das variáveis consideradas; e a subseção de otimização de variáveis controláveis examina o impacto das recomendações geradas, destacando seu potencial para apoiar ajustes práticos nos treinos.

V.1 Resultados da Pesquisa

V.1.1 Base Questionário

A amostra deste estudo foi composta por 26 atletas praticantes de corrida (n = 26), em diferentes níveis de maturidade e tempo de prática. Esse número de respondentes, embora restrito, é representativo de um grupo heterogêneo de corredores recreacionais e amadores, permitindo compreender diferentes perfis dentro da modalidade.

Em termos de composição demográfica, 34,6% dos participantes se identificaram com o sexo feminino, enquanto o restante correspondeu ao sexo masculino. A distribuição etária indica predominância de adultos entre 31 e 40 anos (46,2%), seguidos por indivíduos de 21 a 30 anos (38,5%). As demais faixas etárias apresentaram participação reduzida, sendo 7,7% entre 41 e 50 anos, e 3,8% tanto entre 51 e 60 quanto entre 61 e 64 anos. Não houve respondentes com menos de 21 anos.

Esse panorama etário reforça que a amostra é composta majoritariamente por adultos jovens e de meia-idade, grupo frequentemente associado à prática regular de atividades físicas com objetivos de saúde e desempenho recreativo. A média de peso dos corredores foi de 76,4 kg, e a média de altura, de 1,70 m, valores compatíveis com o perfil antropométrico esperado para adultos fisicamente ativos.

No que se refere à ocupação, observa-se predominância de participantes empregados (80,8%), enquanto 11,5% declararam-se estudantes, 3,8% desempregados e outros 3,8% aposentados. Tal distribuição evidencia que a maioria dos respondentes concilia a prática esportiva com rotinas profissionais, o que pode impactar na disponibilidade de tempo para treinos e na estruturação de planos de treinamento. Além disso, 34,6% dos corredores relataram já ter sofrido algum tipo de lesão ao longo da vida esportiva, indicando que, mesmo entre praticantes recreacionais, a ocorrência de lesões é um fator relevante e recorrente.

Quanto à experiência na modalidade, a maior parte dos participantes (61,5%) afirmou correr há períodos entre 37 e 193 meses, caracterizando um grupo com experiência consolidada na prática. Outros 19,2% acumulam entre 13 e 18 meses de prática, enquanto 15,4% possuem entre 7 e 12 meses, e apenas 3,8% estão iniciando, com menos de 6 meses de experiência. Não foram registrados respondentes com períodos intermediários entre 19 e 36 meses. Esse predomínio de corredores experientes sugere uma amostra composta por indivíduos com hábitos esportivos bem estabelecidos e, possivelmente, maior consciência sobre aspectos de treinamento e recuperação.

A análise comparativa entre os grupos de corredores que já tiveram lesão e aqueles que nunca relataram histórico de lesão revela uma tendência relevante. Observa-se que os participantes com histórico de lesão possuem, em média, maior tempo de prática na modalidade (95 meses) em comparação aos não lesionados (56 meses). Esse resultado sugere uma relação positiva entre a experiência acumulada em corrida e a probabilidade de ocorrência de lesões, possivelmente em decorrência da maior exposição à carga de treino ao longo do tempo. O *boxplot* apresentado a seguir na Figura V.1 ilustra essa diferença, evidenciando uma distribuição mais ampla dos meses de prática entre os corredores lesionados, o que reforça a hipótese de que o tempo de envolvimento com a atividade está associado ao aumento do risco de lesão, ainda que esse risco possa também estar relacionado a fatores como intensidade, volume de treino e histórico individual de recuperação.

A frequência semanal de prática também reflete um comprometimento consistente com a modalidade: 61,5% dos corredores afirmaram treinar de duas a três vezes por semana, e 26,9% relataram correr mais de três vezes por semana. Apenas 11,5% limitam-se a uma sessão semanal. Em relação ao tipo de corrida preferido, a maioria (57,7%) demonstrou preferência por provas de curta distância (até 10 km), enquanto 42,3% afirmaram optar por percursos de longa distância (acima de 10 km). Esse equilíbrio relativo entre curtas e longas distâncias sugere diversidade de objetivos e níveis de preparo físico entre os respondentes.

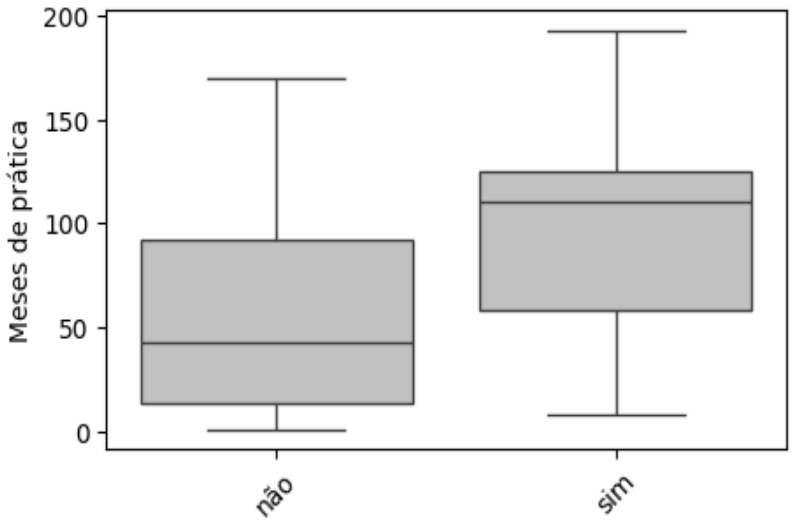


Figura V.1: Distribuição dos meses de prática dos lesionados e não lesionados

No tocante ao acompanhamento profissional, 57,7% dos atletas declararam não contar com assistência especializada, 38,5% relataram treinar sob orientação profissional, e 3,8% indicaram já ter recebido acompanhamento, embora não o façam atualmente. Esse dado é relevante, pois o suporte técnico é um fator associado à prevenção de lesões e à otimização do desempenho, especialmente entre corredores que mantêm frequência elevada de treinos.

O ambiente de prática preferido reforça a predominância de corredores que optam por espaços abertos: 88,5% preferem correr em vias públicas, 7,7% utilizam a esteira, e 3,8% realizam treinos em trilhas. Essa predominância de corridas ao ar livre está em consonância com estudos que apontam o ambiente urbano como o principal espaço de prática para corredores amadores, devido à acessibilidade e à conveniência.

Por fim, observou-se que a maioria dos participantes também realiza outras modalidades de exercício físico. Metade dos respondentes (50,0%) pratica outras atividades de duas a três vezes por semana, enquanto 46,2% o fazem mais de três vezes por semana e 3,8% apenas uma vez. Entre essas atividades complementares, a musculação se destaca como a mais comum, sendo praticada por 88,5% dos atletas. Tal dado reforça a importância atribuída ao fortalecimento muscular como estratégia de suporte ao desempenho e à prevenção de lesões no contexto da corrida.

De modo geral, o perfil dos respondentes revela um grupo de corredores predominantemente adultos e engajados, com elevada frequência de prática e hábitos complementares de exercício físico. Esses elementos contribuem para contextualizar as análises subsequentes deste estudo, especialmente aquelas relacionadas à carga de treino e desempenho entre corredores recreacionais. No mais, todas as estatísticas descritivas apresentadas nesta seção foram calculadas com base nas informações declaradas pelos próprios atletas nos questionários aplicados.

V.1.2 Base Atividades

A análise da base de dados proveniente do compartilhamento de atividades esportivas revelou um conjunto expressivo de informações sobre o comportamento de corrida dos participantes. Dos 26 atletas que responderam ao questionário, 21 consentiram em disponibilizar seus registros do aplicativo Strava. Após os procedimentos de limpeza e padronização dos dados, como já descritos nas etapas metodológicas, foram analisadas 1.599 sessões de treino, todas contendo as principais métricas de desempenho, como velocidade média, distância e duração. Esse volume de dados possibilitou uma visão mais precisa e objetiva sobre o desempenho dos corredores em diferentes condições, complementando as informações declaradas no questionário.

No que se refere à relação entre desempenho e ocorrência de lesões, observou-se uma diferença clara nas velocidades médias registradas em períodos com e sem lesão. A velocidade média das corridas realizadas em momentos sem lesão foi de 9,45 km/h, enquanto, durante períodos em que os atletas relataram estar lesionados, a média caiu para 8,75 km/h. Essa redução de desempenho é consistente com o que se observa na literatura esportiva, em que quadros de dor ou limitação física tendem a impactar negativamente a intensidade e a eficiência da corrida. O gráfico na na Figura V.2 correspondente evidencia essa diferença, indicando que a condição de lesão está associada a uma desaceleração perceptível no ritmo de treino.

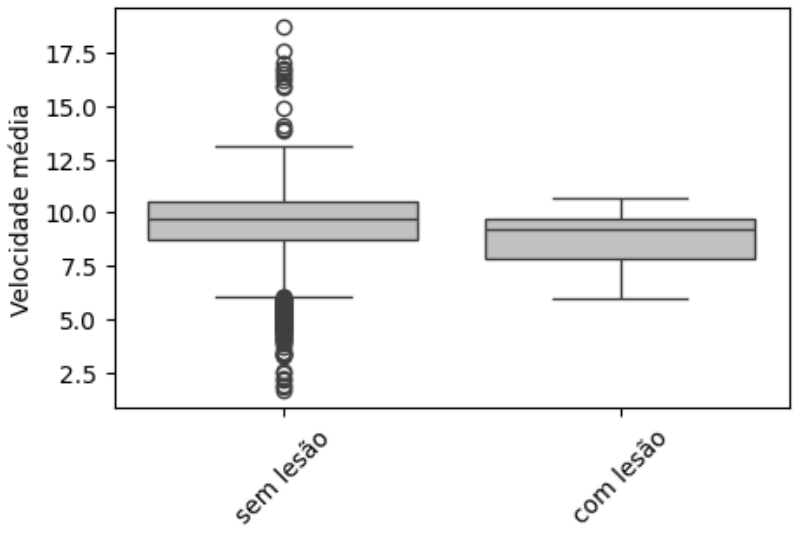


Figura V.2: Distribuição da velocidade média por períodos de lesão

Outro aspecto relevante identificado diz respeito à influência da prática de outros tipos de exercício físico sobre a velocidade média de corrida. Observou-se que atletas que praticam outros exercícios apenas uma vez por semana apresentaram velocidade média de 13,4 km/h, enquanto aqueles que treinam de duas a três vezes por semana registraram média de 9,33 km/h, e os que praticam mais de três vezes por semana obtiveram média de 9,48 km/h. Esse resultado sugere uma

relação inversa entre a frequência de prática de outras modalidades e a velocidade média na corrida.

Uma possível interpretação para esse achado é que indivíduos que dedicam menos tempo a outras atividades físicas podem concentrar maior volume e intensidade de treinos de corrida, o que resultaria em melhor desempenho específico na modalidade. O *boxplot* na Figura V.3 ilustra essa variação, mostrando uma dispersão maior entre os corredores que praticam outras atividades com menor frequência, o que pode refletir diferentes níveis de dedicação e objetivos esportivos individuais.

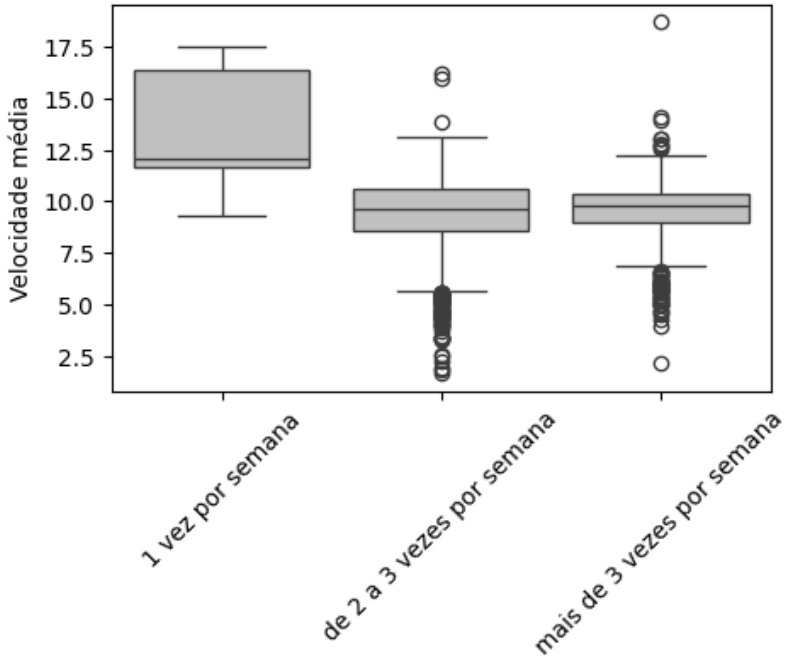


Figura V.3: Distribuição da velocidade média por frequência de prática de outros exercícios

De modo geral, essas análises revelam que tanto a condição física (ausência de lesão) quanto o equilíbrio entre modalidades de treino podem influenciar diretamente o desempenho médio dos corredores. Tais resultados reforçam a importância de compreender as interações entre carga de treino, recuperação e prática complementar, fornecendo evidências quantitativas que podem subsidiar recomendações para uma prática mais segura e eficiente da corrida.

V.2 Resultados da Modelagem

Antes de apresentar os resultados da modelagem, é importante destacar uma limitação central deste estudo relacionada ao tamanho relativamente reduzido da amostra de participantes (n = 26), o que pode afetar a robustez e a estabilidade das análises estatísticas conduzidas. Apesar de a base atividades conter 1.599 observações de treino, que servirá como *input* para o modelo de previsão de desempenho, o pequeno número de unidades independentes aumenta a variabilidade das estimativas dos coeficientes da regressão e eleva o risco de sobreajuste do modelo. Consequentemente, a capacidade preditiva do modelo pode estar superestimada quando aplicada fora da amostra, e a

precisão das inferências sobre o efeito das variáveis explicativas no desempenho dos atletas pode ser limitada.

Além disso, a identificação de *clusters* com um número reduzido de participantes apresenta maior sensibilidade ao ruído amostral, podendo gerar agrupamentos instáveis ou menos representativos dos padrões reais de comportamento de treino. Essa instabilidade dificulta a replicabilidade dos resultados em amostras independentes e restringe a generalização dos perfis identificados. Portanto, embora os resultados forneçam *insights* valiosos para o entendimento das trajetórias de desempenho, recomenda-se cautela na interpretação e destaca-se a necessidade de estudos futuros com amostras maiores para validação e refinamento dos achados.

V.2.1 Fase 1 - Encontrando Perfil dos Corredores

Na etapa de clusterização, buscou-se identificar padrões latentes entre os corredores a partir de suas características físicas e de experiência na modalidade, a partir das informações de idade do atleta, peso, altura e tempo total desde o início das atividades de corrida. Conforme mencionado anteriormente, foi aplicada a técnica PCA com o objetivo de reduzir a dimensionalidade dos dados e capturar 80% da variabilidade total. A análise resultou na identificação de duas componentes principais, responsáveis por representar de maneira adequada a estrutura de variação existente no conjunto de dados originais. O gráfico na Figura V.4 de variância explicada demonstra que a adoção de duas dimensões é suficiente para preservar a maior parte da informação relevante, garantindo um equilíbrio entre simplicidade e representatividade dos dados.

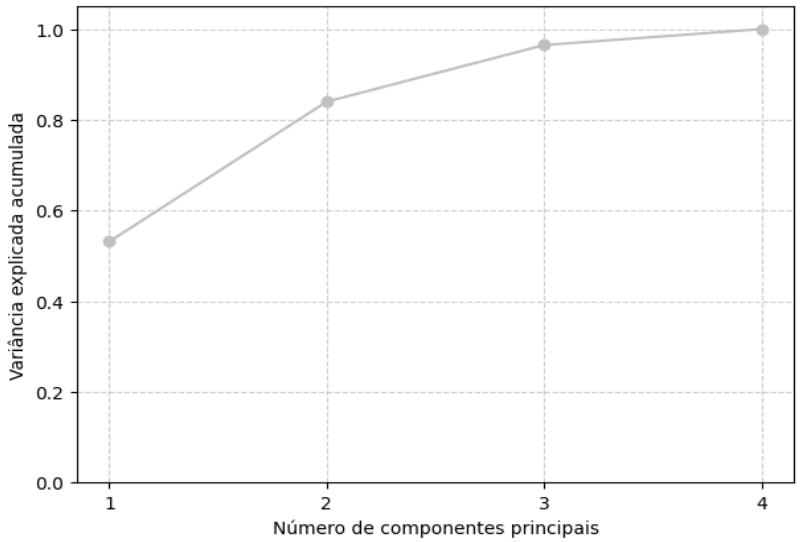


Figura V.4: Variabilidade capturada por componente

Na sequência, diferentes algoritmos de agrupamento foram aplicados sobre o espaço transformado pelas componentes principais. O método *K-means* indicou a formação de quatro *clusters* como ponto ótimo, conforme o método do cotovelo; o *Agglomerative Clustering* (hierárquico) sugeriu a existência

de três grupos distintos, enquanto o DBSCAN identificou apenas um *cluster* relevante, classificando os demais pontos como *outliers*.

A comparação dos índices de silhueta indicou a superioridade do modelo aglomerativo (0,417) em relação aos métodos *K-means* (0,364) e DBSCAN (0,382), evidenciando melhor equilíbrio entre compacidade e separação dos grupos. Como critério complementar, avaliou-se também o índice de Davies–Bouldin, cujos valores foram 0,775 para o *K-means*, 0,765 para o modelo aglomerativo e 0,651 para o DBSCAN. Embora o DBSCAN tenha apresentado o menor valor nessa métrica, a diferença entre *K-means* e aglomerativo foi pequena, enquanto, na silhueta, a vantagem do modelo aglomerativo foi mais expressiva. Dessa forma, considerando conjuntamente as métricas e atribuindo maior peso à capacidade de separação evidenciada pela silhueta, o modelo aglomerativo foi selecionado como abordagem final. O dendrograma resultante dessa análise hierárquica foi utilizado para visualizar as relações de proximidade entre os atletas, evidenciando a estrutura natural dos grupos identificados, como mostra a Figura V.5.

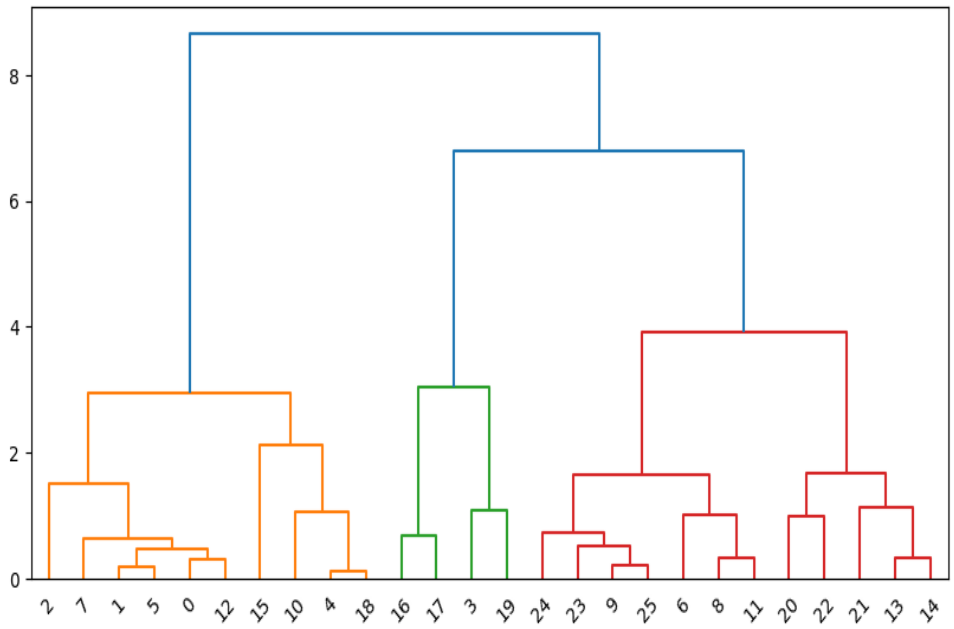


Figura V.5: Dendograma - Agglomerative Clustering

A interpretação dos agrupamentos revelou perfis bem definidos entre os corredores. O *cluster* 0, denominado "Experientes Jovens", reúne atletas com idade média de 30 anos, em sua maioria sem histórico de lesão (apenas 17%) e com experiência média de 71 meses na corrida. O *cluster* 1, classificado como "Pouco Experientes", também é composto majoritariamente por jovens (média de 30 anos), porém com menor tempo de prática (39 meses) e uma proporção de 30% de indivíduos que já tiveram lesão. Por fim, o *cluster* 2, designado "Experientes Veteranos", agrupa os corredores mais velhos (média de 53 anos), todos com histórico de lesão (100%) e elevado tempo de experiência na modalidade (142 meses). Esses resultados demonstram uma clara segmentação entre jovens inici-

antes e corredores mais maduros e experientes, destacando o papel da experiência na diferenciação dos perfis.

Com o intuito de verificar se essas diferenças eram estatisticamente significativas, foi aplicado o teste não paramétrico de *Mann–Whitney*[1], comparando o tempo de experiência declarado (em meses de prática) entre os grupos mais experientes (*clusters* 0 e 2 combinados) e o grupo menos experiente (*cluster* 1). O resultado apresentou relevância estatística ($p < 0{,}05$), confirmando diferença entre os grupos e corroborando a hipótese de que os corredores mais experientes possuem trajetória substancialmente mais longa na modalidade.

Além disso, a partir das evidências apresentadas no estudo de Tito et al. [2024], que demonstrou que atletas mais experientes apresentam menor número de pontos de mudança em seus treinos, buscou-se verificar se essa estabilidade também se refletia na consistência do desempenho entre os *cluster* identificados. Para isso, foi realizado um novo teste de hipótese comparando os desvios padrão das velocidades entre os grupos de corredores experientes (*clusters* 0 e 2 combinados) e o grupo menos experiente (*cluster* 1). Os resultados mostraram que os corredores menos experientes apresentaram maior variabilidade em seus treinos (desvio médio na velocidade de 2,59) em comparação aos mais experientes (2,34), diferença também significativa.

Essa constatação está em consonância com a literatura, reforçando a ideia de que a experiência tende a reduzir a flutuação de desempenho e aumentar a consistência dos treinos. O gráfico na Figura V.6 ilustra essa diferença de dispersão entre os grupos, evidenciando que os corredores experientes mantêm padrões mais estáveis de velocidade ao longo de suas atividades, o que pode refletir maior controle de esforço, melhor percepção corporal e estratégias mais maduras de gestão do ritmo durante os treinos.

V.2.2 Fase 2 - Prevendo Desempenho dos Corredores

Como mencionado na seção de Metodologia, o modelo de previsão da velocidade média do treino foi desenvolvido a partir de um conjunto abrangente de variáveis explicativas. Após a divisão da base de dados em conjuntos de treino e teste e a padronização das variáveis explicativas, foi treinado inicialmente um modelo geral, de caráter mais abrangente, utilizando toda a base de atletas, no qual foram avaliados dois algoritmos de regressão: a Regressão Linear e o XGBoost. O modelo de Regressão Linear apresentou R^2 de 0,31 e RMSE de 1,36 no conjunto de teste; contudo, ao aplicar a validação cruzada por meio do método *GroupKFold*, o RMSE médio aumentou para 1,88, indicando limitação na capacidade de generalização. Em contrapartida, o modelo XGBoost obteve desempenho superior, com R^2 de 0,53 e RMSE de 1,12 no conjunto de teste, e apresentou RMSE médio de 1,87 após a validação cruzada com *GroupKFold*. Diante desses resultados, o modelo

[1]Disponível em: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html`. Último acesso em: 10 de novembro de 2025.

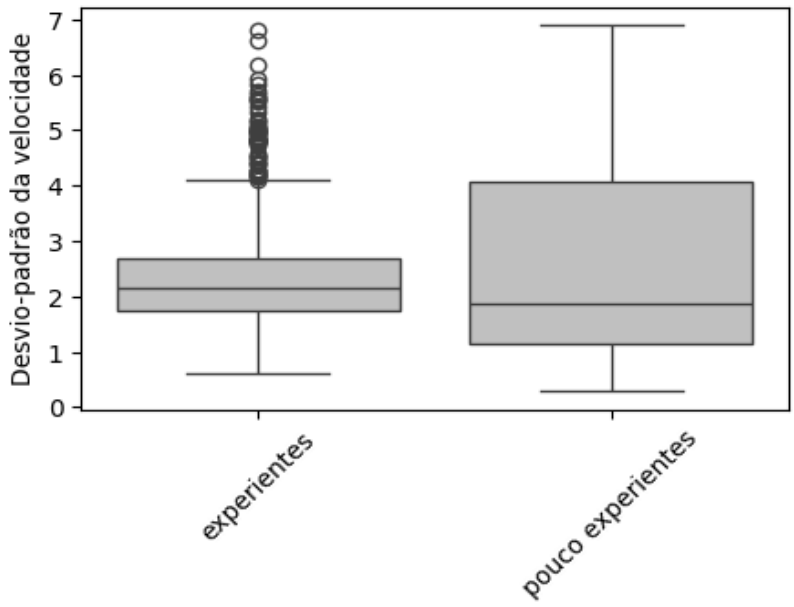


Figura V.6: Distribuição do desvio padrão da velocidade

XGBoost foi selecionado como o modelo final, por apresentar melhor desempenho preditivo e maior robustez frente à validação cruzada.

Complementarmente, foi elaborado um gráfico de SHAP *values* com o objetivo de ilustrar a contribuição individual de cada variável na previsão da velocidade média a partir do modelo selecionado, conforme apresentado na Figura V.7. Observa-se que as variáveis relacionadas ao próprio desempenho do treino exercem maior influência sobre as predições, com destaque para a velocidade mínima e máxima do treino, que apresentaram os maiores valores médios de contribuição, seguidas pela distância percorrida e pelo desvio-padrão da velocidade, indicando a relevância tanto da intensidade quanto da variabilidade do esforço. Além disso, fatores contextuais e históricos, como a temperatura durante o treino e as velocidades médias de sessões anteriores, também mostraram impacto relevante, sugerindo que o modelo incorpora efeitos de carga acumulada e condições ambientais na estimativa da performance. Variáveis de caráter mais estrutural, como o tempo total desde o início das atividades de corrida e a presença de auxílio profissional, apresentaram contribuição moderada, reforçando o papel da experiência e do acompanhamento técnico no desempenho médio. De maneira geral, a distribuição das importâncias indica que o modelo captura padrões amplos e consistentes entre os atletas, priorizando características diretamente observáveis do treino e do histórico recente, o que é coerente com sua proposta mais generalista e com a necessidade de robustez em um cenário com elevada heterogeneidade entre indivíduos.

Em complemento à análise de interpretabilidade, foi conduzida uma análise de sensibilidade com o objetivo de avaliar o impacto marginal das principais variáveis sobre a velocidade média prevista. Observou-se que aumentos nas variáveis diretamente associadas ao desempenho do treino tendem a

gerar impacto positivo na predição, com destaque para a velocidade mínima e a velocidade máxima, que apresentaram os maiores efeitos incrementais. A distância percorrida e as velocidades médias de treinos anteriores também exibiram influência positiva, reforçando o papel da carga recente e do histórico de desempenho na estimativa da performance. Por outro lado, algumas variáveis apresentaram impacto negativo sobre a velocidade média prevista, como a temperatura média do treino e o desvio-padrão da velocidade, sugerindo que condições ambientais menos favoráveis e maior variabilidade de ritmo estão associadas a reduções no desempenho estimado. Variáveis de caráter mais estrutural, como o tempo total de prática de corrida, mostraram efeito negativo moderado, enquanto a presença de acompanhamento profissional apresentou impacto praticamente nulo. De forma geral, os resultados da análise de sensibilidade corroboram os achados dos SHAP *values*, indicando que o modelo não apenas identifica as variáveis mais relevantes, mas também captura de maneira consistente a direção de seus efeitos sobre a velocidade média.

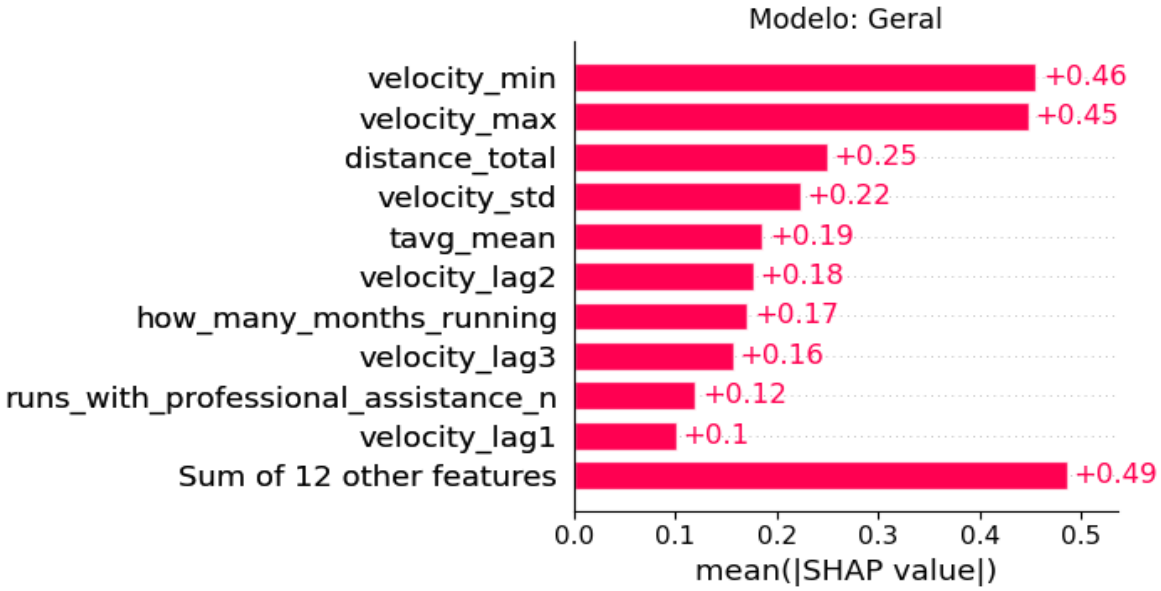


Figura V.7: SHAP *values* - Contribuição das variáveis (Geral)

Em seguida, buscando uma compreensão mais detalhada e aderente às características dos diferentes perfis de corredores, foram estimados modelos, separados por *cluster*, definidos a partir do processo de segmentação descrito anteriormente. Essa abordagem permitiu capturar diferenças estruturais no comportamento de cada grupo.

No *cluster* 0 ("Experientes Jovens"), também foram avaliados os modelos de Regressão Linear e XGBoost para a previsão da velocidade média, sendo o XGBoost selecionado por apresentar melhor desempenho. O modelo de Regressão Linear obteve R^2 de 0,40 e RMSE de 1,71 no conjunto de teste; entretanto, após a aplicação da validação cruzada por meio do método *GroupKFold*, o RMSE médio aumentou para 2,54, indicando menor robustez. Já o modelo XGBoost apresentou R^2 de 0,49 e RMSE de 1,57, com RMSE médio de 1,99 na validação cruzada, sendo, portanto, escolhido

como o modelo final para esse grupo.

A análise de interpretabilidade por meio dos SHAP *values*, apresentada na Figura V.8, indica que, no *cluster* 0, variáveis diretamente associadas à intensidade do treino são as mais influentes nas predições do modelo, com destaque para a velocidade máxima e a velocidade mínima do treino. Esses resultados reforçam a importância tanto dos picos de esforço quanto do nível mínimo de intensidade sustentado ao longo da sessão. Além disso, características individuais como a altura e a idade do atleta no momento do treino apresentaram contribuição relevante, sugerindo que fatores antropométricos e demográficos ainda desempenham papel significativo mesmo em atletas experientes. Variáveis relacionadas à carga recente e ao contexto do treino, como as velocidades médias dos treinos anteriores, a distância percorrida, o desvio-padrão da velocidade e a temperatura, também contribuíram de forma moderada. Em conjunto, os resultados indicam que o modelo prioriza predominantemente aspectos ligados à intensidade do treino, modulados por características individuais e pelo histórico recente de carga nesse grupo.

Em complemento à análise de interpretabilidade, a análise de sensibilidade evidenciou que, para atletas jovens e experientes, aumentos nas variáveis associadas à intensidade do treino exercem impacto positivo direto sobre a velocidade média prevista, com destaque para a velocidade mínima e a velocidade máxima, que apresentaram os maiores efeitos incrementais. As velocidades médias de treinos anteriores e a distância percorrida também mostraram influência positiva, reforçando a relevância da carga recente e da continuidade do estímulo no desempenho estimado. A idade do atleta apresentou efeito positivo moderado, sugerindo que, dentro desse grupo específico, a maturidade esportiva contribui de forma favorável para a performance prevista. Por outro lado, algumas variáveis apresentaram impacto negativo quando aumentadas, como o desvio-padrão da velocidade, indicando que maior variabilidade de ritmo tende a reduzir a velocidade média estimada. A altura, a temperatura média do treino e a frequência de prática de outros exercícios também exibiram efeitos negativos, ainda que moderados, sugerindo que fatores antropométricos, ambientais e de carga concorrente podem atenuar o desempenho previsto. Em conjunto, os resultados da análise de sensibilidade são consistentes com os SHAP *values*, reforçando a robustez do modelo ao capturar tanto a relevância quanto a direção dos efeitos das principais variáveis nesse grupo.

No *cluster* 1 ("Pouco Experientes"), observou-se um ganho expressivo de desempenho ao empregar o modelo XGBoost em comparação à Regressão Linear. O modelo linear apresentou R^2 de 0,37 e RMSE de 1,26 no conjunto de teste; contudo, ao aplicar a validação cruzada com o método *GroupKFold*, o RMSE médio elevou-se significativamente para 8,34, evidenciando forte instabilidade e baixa capacidade de generalização nesse grupo. Em contraste, o modelo XGBoost obteve R^2 de 0,44 e RMSE de 1,19 no conjunto de teste, mantendo desempenho consistente na validação cruzada, com RMSE médio de 1,29, o que motivou sua escolha como modelo final para esse *cluster*.

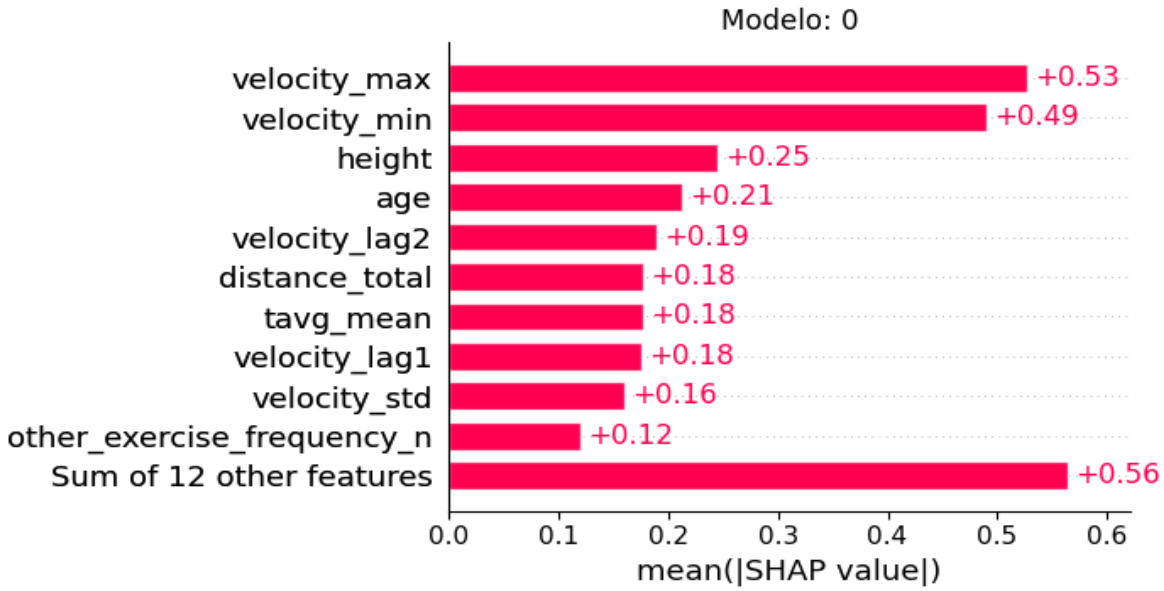


Figura V.8: SHAP *values* - Contribuição das variáveis (*Cluster* 0)

A análise dos SHAP *values*, apresentada na Figura V.9, indica que, entre corredores novatos, a variabilidade do ritmo exerce papel central na previsão da velocidade média, com destaque para o desvio-padrão da velocidade do treino como a variável de maior contribuição, o que reflete a menor estabilidade de intensidade característica desse perfil. Esse resultado mostra-se particularmente relevante por estar em consonância com os achados de Tito et al. [2024], mesmo tendo sido obtido por meio de uma abordagem metodológica distinta. No referido estudo, a análise foi conduzida a partir de séries temporais e detecção de pontos de mudança nos treinos, e a classificação do nível de experiência dos corredores baseou-se no tempo de prática, e não em técnicas de clusterização. A convergência dos resultados, apesar dessas diferenças de modelagem e de definição dos perfis, reforça a robustez da evidência de que a variabilidade do ritmo é um fator central na caracterização e no desempenho de corredores menos experientes.

Adicionalmente, a velocidade máxima, a distância percorrida e a velocidade mínima do treino apresentaram influência relevante, sugerindo que tanto picos de esforço quanto a capacidade de sustentar o ritmo impactam fortemente o desempenho estimado. Variáveis relacionadas ao histórico recente, como as velocidades médias de treinos anteriores, também contribuíram de forma significativa, indicando que o modelo capta adaptações iniciais ao treinamento. Por fim, fatores individuais e contextuais, como a idade do atleta, a frequência de prática de outros exercícios e a temperatura durante o treino, apresentaram menor, porém não desprezível, contribuição, reforçando que, para atletas pouco experientes, o desempenho é mais sensível a flutuações de ritmo, carga e contexto, o que justifica a superioridade da modelagem de relações não lineares nesse grupo.

Complementando a análise de interpretabilidade, a análise de sensibilidade mostra que, entre corredores menos experientes, incrementos nas variáveis associadas à intensidade do treino produzem

efeito positivo na velocidade média prevista, com destaque para a velocidade mínima e a velocidade máxima. As velocidades médias de treinos anteriores e a distância percorrida também apresentaram impacto positivo relevante, indicando que o modelo é sensível ao histórico recente e à progressão inicial da carga de treino. Em contraste, o desvio-padrão da velocidade exibiu impacto negativo pronunciado, sugerindo que maior instabilidade de ritmo está associada a reduções no desempenho estimado. A idade do atleta e a temperatura média do treino também mostraram efeitos negativos, embora de menor magnitude. De forma geral, os resultados também são consistentes com os SHAP *values*, reforçando que, nesse grupo, a estabilidade do ritmo e a adaptação ao treinamento são determinantes centrais da performance prevista.

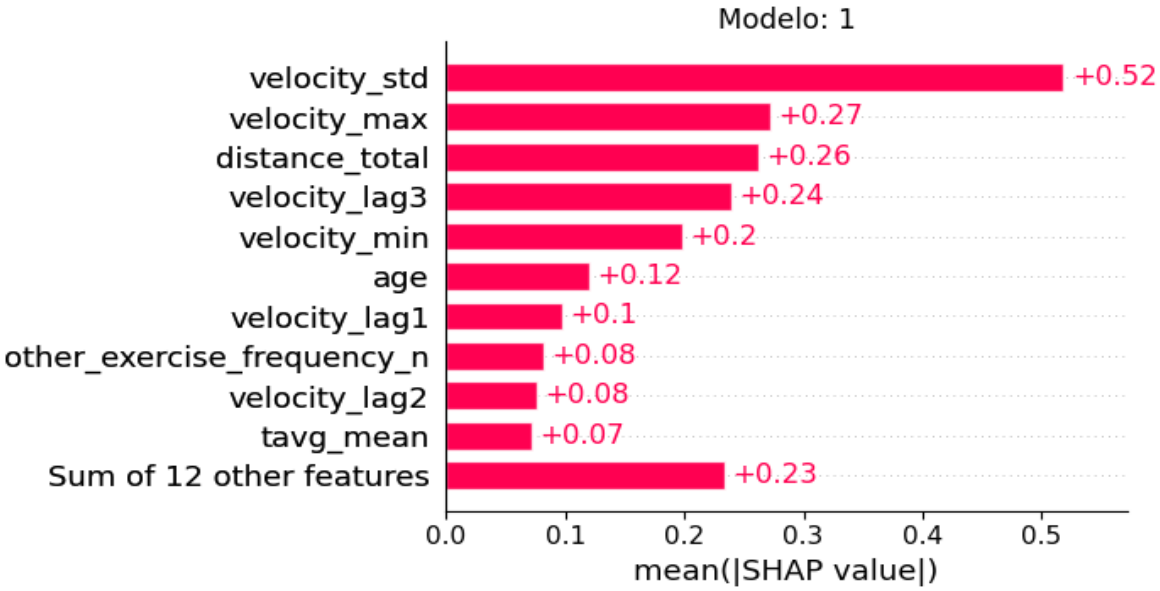


Figura V.9: SHAP *values* - Contribuição das variáveis (*Cluster* 1)

Por fim, no *cluster* 2 (“Experientes Veteranos”), observou-se novamente a superioridade do modelo XGBoost em relação à Regressão Linear, especialmente no que se refere à robustez sob validação cruzada. O modelo de Regressão Linear apresentou R^2 de 0,47 e RMSE de 0,73 no conjunto de teste; entretanto, ao aplicar a validação cruzada por meio do método *GroupKFold*, o RMSE médio elevou-se para 5,59, evidenciando limitação na capacidade de generalização mesmo em um grupo mais homogêneo. Em contraste, o modelo XGBoost obteve R^2 de 0,51 e RMSE de 0,70 no conjunto de teste, mantendo desempenho consistente na validação cruzada, com RMSE médio de 1,08, o que motivou sua escolha como modelo final para esse *cluster*.

A análise dos SHAP *values*, apresentada na Figura V.10, indica que, para atletas experientes e mais velhos, o histórico recente de desempenho exerce papel central nas predições, com destaque para a velocidade média do treino anterior como a variável de maior contribuição. O desvio-padrão da velocidade do treino e a distância percorrida também apresentaram influência relevante, sugerindo que a consistência do ritmo e o volume do treino são determinantes importantes nesse perfil.

Variáveis associadas à recuperação, como os dias desde o último treino, mostraram contribuição moderada, reforçando a importância da gestão da carga e dos intervalos de descanso para atletas veteranos. Em contrapartida, indicadores de intensidade pontual, como as velocidades máxima e mínima do treino, bem como variáveis de organização do ciclo, apresentaram menor impacto relativo. Em conjunto, os resultados sugerem que, nesse grupo, o desempenho é mais fortemente ancorado no histórico recente e na estabilidade do treinamento ao longo do tempo.

Em complemento à análise de interpretabilidade, a análise de sensibilidade indicaram que, para atletas experientes e mais velhos, aumentos em variáveis associadas à estabilidade e ao histórico recente de desempenho tendem a exercer impacto positivo sobre a velocidade média prevista. Nesse contexto, o desvio-padrão da velocidade e as velocidades médias de treinos anteriores apresentaram efeitos incrementais moderados, juntamente com as velocidades mínima e máxima do treino, sugerindo que tanto a consistência do ritmo quanto a manutenção da intensidade contribuem para o desempenho estimado. Por outro lado, os dias desde o último treino apresentaram impacto negativo, indicando que intervalos mais longos entre sessões estão associados a reduções na velocidade média prevista. Variáveis relacionadas ao volume e à organização do ciclo de treinamento, como a distância percorrida e o número de treinos no mês, também exibiram efeitos negativos, ainda que de menor magnitude. De modo geral, os resultados aqui também são coerentes com os SHAP *values*, reforçando que, nesse grupo, o desempenho é fortemente condicionado pela continuidade do treinamento e pela gestão adequada dos períodos de recuperação.

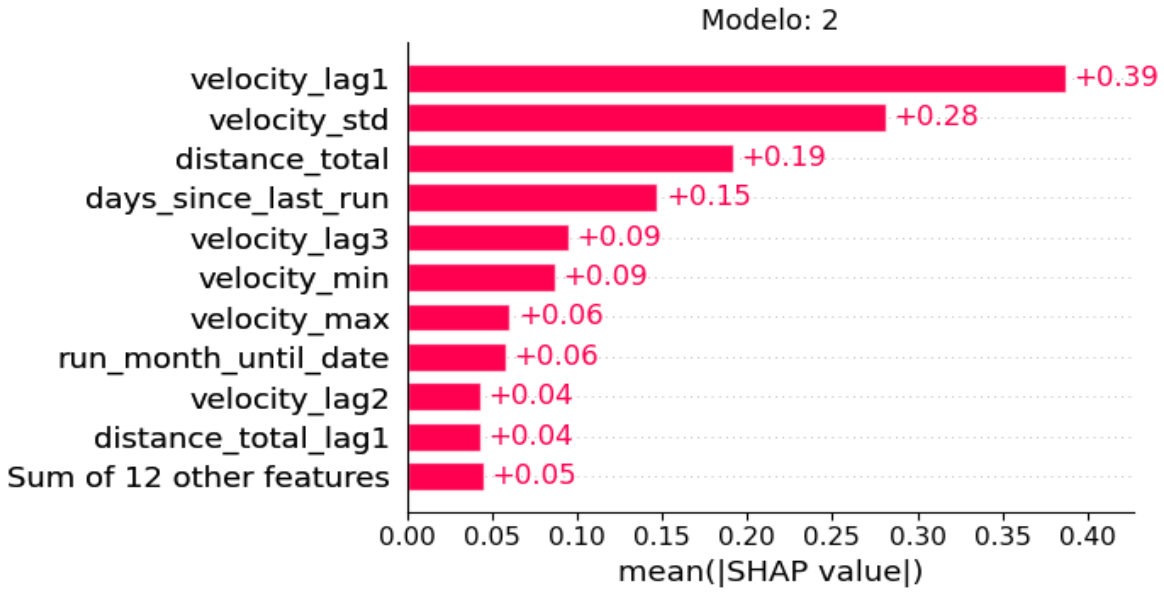


Figura V.10: SHAP *values* - Contribuição das variáveis (*Cluster* 2)

De forma geral, os resultados evidenciam que o desempenho dos modelos e os fatores que influenciam a previsão da velocidade média variam de maneira consistente entre os diferentes perfis de atletas. Enquanto o modelo generalista e os *clusters* de corredores experientes tendem a ser mais

influenciados por métricas diretas de intensidade e histórico recente de desempenho, os corredores pouco experientes apresentam maior sensibilidade à variabilidade do ritmo e a flutuações de carga, o que reforça a importância de modelos capazes de capturar relações não lineares. Em todos os cenários, o XGBoost demonstrou maior robustez e capacidade de generalização, especialmente sob validação cruzada com *GroupKFold*, mostrando-se mais adequado para lidar com a heterogeneidade dos dados e com as diferentes dinâmicas de treinamento associadas a níveis distintos de experiência e idade.

V.2.3 Fase 3 - Recomendação Personalizada por Perfil

A etapa de otimização teve como objetivo identificar as combinações ideais das variáveis controláveis que maximizam a velocidade média prevista dos treinos, conforme os modelos de regressão previamente treinados, e para isso, foi empregada a função `minimize`, disponível na biblioteca SciPy. As condições de restrição, limites de variáveis (*bounds*) e valores iniciais (*initial points*) foram definidos conforme descrito na seção de Metodologia, garantindo que as soluções obtidas respeitassem os intervalos plausíveis e as restrições lógicas estabelecidas para o problema.

O procedimento foi aplicado tanto ao modelo geral, construído com a base total de treinos, quanto aos modelos segmentados por agrupamento de atletas, conforme definido na fase de clusterização. Essa abordagem permitiu capturar padrões específicos de otimização entre grupos de atletas com diferentes perfis de experiência, idade e histórico de lesão.

As variáveis selecionadas para o processo de otimização foram aquelas consideradas controláveis, ou seja, atributos sobre os quais o atleta possui influência direta, tornando-as adequadas para fins de recomendação. Entre elas estão: velocidade máxima e mínima do treino, distância total percorrida, número de treinos realizados na semana, frequência de prática de outros exercícios, prática de musculação (variável booleana, sendo 1 para "sim" e 0 para "não") e tipo de treino (variável booleana, sendo 1 para "intervalado" e 0 para "não intervalado"). Adicionalmente, a escolha dessas variáveis foi respaldada pela análise dos valores de SHAP, que indicaram sua relevância na explicação das previsões do modelo, contribuindo para embasar e refinar uma seleção que, a priori, poderia ser considerada arbitrária. Dessa forma, o resultado da otimização fornece não apenas uma previsão teórica de desempenho, mas também parâmetros práticos que podem orientar o planejamento individualizado de treinos.

Os resultados da otimização global (Tabela V.1) indicaram uma velocidade média prevista de 10,64 km/h, obtida a partir de uma combinação de parâmetros que refletem um perfil moderado e equilibrado de treinamento. O valor ótimo de distância (6,95 km) associado a uma frequência semanal de dois treinos sugere que, no contexto de um modelo geral, a maximização do desempenho não está vinculada a volumes excessivos, mas a uma carga de treino sustentável. Observa-se também

a relevância da prática de musculação, indicada pelo valor ótimo igual a 1, enquanto o tipo de treino não intervalado (valor 0) aparece como a configuração mais favorável nesse cenário agregado. Esses resultados reforçam a ideia de que, ao considerar atletas de forma conjunta e sem distinção de perfis, o modelo tende a privilegiar estratégias de treino mais conservadoras e consistentes, nas quais o equilíbrio entre estímulo físico e recuperação desempenha papel central na otimização do desempenho médio previsto.

A análise de sensibilidade da etapa de otimização indicou que a velocidade média prevista é mais sensível a variações na variável velocidade mínima do treino, que apresentou o maior impacto positivo ao ser incrementada em um desvio-padrão, resultando em aumento de aproximadamente 0,18 km/h na predição. Em menor magnitude, número de treinos dentro da semana também mostrou influência positiva, sugerindo que maior consistência semanal de treinos contribui para ganhos adicionais de desempenho. As demais variáveis analisadas não apresentaram variação na velocidade média prevista quando perturbadas individualmente, indicando baixa sensibilidade local do modelo a esses fatores no entorno da solução ótima.

Tabela V.1: Resultado da otimização da função objetivo (velocidade média prevista) - Geral

Variável	**Valor ótimo**
Velocidade máxima do treino (km/h)	15.27
Velocidade mínima do treino (km/h)	0.00
Distância (km)	6.95
Número de treinos dentro da semana	2.00
Frequência de prática de outros exercícios na semana	2.00
Prática de musculação	1.00
Tipo de treino	0.00
Velocidade média prevista (função objetivo)	**10.64**

Ao analisar as otimizações por *cluster* (Tabelas V.2, V.3 e V.4), observam-se padrões alinhados às características esperadas de cada grupo de atletas. O *cluster* 0 ("Experientes Jovens") apresentou o maior valor ótimo de velocidade média prevista (12,11 km/h), associado a uma velocidade máxima relativamente mais elevada (16,76 km/h) e à maior frequência de prática de outros exercícios na semana. Esse resultado sugere que, para atletas mais jovens e experientes, a diversidade de estímulos físicos, combinada a uma intensidade moderada, desempenha papel relevante na maximização do desempenho.

Nesse grupo, a análise de sensibilidade da otimização mostrou que a velocidade média prevista apresenta baixa sensibilidade às variáveis analisadas, com efeitos nulos para a maioria delas quando perturbadas em um desvio-padrão. Destaca-se que incrementos na velocidade mínima do treino e no número de treinos dentro da semana resultam em variações negativas na velocidade média prevista, ainda que de pequena magnitude, indicando que, no entorno da solução ótima, aumentos adicionais

nessas variáveis podem não ser benéficos para esse perfil específico. Esse comportamento sugere que, para atletas mais experientes, o modelo identifica um ponto de equilíbrio mais restritivo, no qual ganhos adicionais de intensidade ou volume tendem a apresentar retornos marginais ou até ligeiramente adversos sobre a velocidade média.

O *cluster* 1 ("Pouco Experientes") apresentou uma velocidade média ótima intermediária (11,65 km/h), obtida com valores semelhantes de velocidade máxima, distância ligeiramente superior (7,97 km) e frequência moderada de treinos e atividades complementares. Esse padrão indica que, para atletas em estágio inicial, a construção gradual de volume e regularidade tende a ser mais determinante para o desempenho do que estratégias baseadas em intensificação precoce.

Para esse grupo de atletas, a análise de sensibilidade da otimização indicou que a velocidade média prevista é predominantemente sensível à variável de frequência de prática de outros exercícios na semana, cujo aumento em um desvio-padrão resulta em uma redução expressiva da velocidade média prevista. Esse resultado sugere que a maior frequência de atividades complementares pode gerar efeito de sobrecarga ou competição com os treinos de corrida, impactando negativamente o desempenho médio desse perfil. Em contraste, a velocidade mínima do treino apresentou um efeito positivo, embora modesto, indicando que pequenos aumentos na intensidade mínima dos treinos podem contribuir para ganhos na velocidade média. As demais variáveis não apresentam variação relevante, reforçando que, para atletas menos experientes, a qualidade e a distribuição do estímulo parecem ser mais determinantes.

Por fim, o *cluster* 2 ("Experientes Veteranos") exibiu a menor velocidade média prevista (7,62 km/h), compatível com o perfil de atletas mais velhos. Os valores ótimos apontam para menor frequência semanal de treinos e uma combinação conservadora de intensidade e volume, reforçando uma estratégia focada em sustentabilidade, controle de carga e prevenção de sobrecarga, em detrimento da maximização da velocidade. De forma geral, os resultados evidenciam que a segmentação por *cluster* permite capturar diferenças relevantes nas estratégias ótimas de treino, impossíveis de serem observadas em um modelo agregado.

Aqui, a análise de sensibilidade da otimização mostrou que a velocidade média prevista apresenta baixa sensibilidade às variáveis analisadas, indicando um cenário de maior estabilidade do desempenho. A única variável com efeito observável foi a velocidade mínima do treino, cujo aumento em um desvio-padrão provoca uma leve redução na velocidade média prevista, sugerindo que elevar excessivamente a intensidade mínima dos treinos pode ser contraproducente para esse perfil. As demais variáveis não apresentam impacto relevante, o que reforça a ideia de que, entre atletas veteranos experientes, o desempenho já se encontra próximo de um patamar consolidado, sendo menos responsivo a ajustes marginais nas características do treinamento.

Tabela V.2: Resultado da otimização da função objetivo (velocidade média prevista) - *Cluster* 0

Variável	**Valor ótimo**
Velocidade máxima do treino (km/h)	16.76
Velocidade mínima do treino (km/h)	0.00
Distância (km)	6.17
Número de treinos dentro da semana	2.00
Frequência de prática de outros exercícios na semana	4.00
Prática de musculação	1.00
Tipo de treino	0.00
Velocidade média prevista (função objetivo)	**12.11**

Tabela V.3: Resultado da otimização da função objetivo (velocidade média prevista) - *Cluster* 1

Variável	**Valor ótimo**
Velocidade máxima do treino (km/h)	14.49
Velocidade mínima do treino (km/h)	0.00
Distância (km)	7.97
Número de treinos dentro da semana	2.00
Frequência de prática de outros exercícios na semana	2.00
Prática de musculação	1.00
Tipo de treino	0.00
Velocidade média prevista (função objetivo)	**11.65**

Tabela V.4: Resultado da otimização da função objetivo (velocidade média prevista) - *Cluster* 2

Variável	**Valor ótimo**
Velocidade máxima do treino (km/h)	15.08
Velocidade mínima do treino (km/h)	0.23
Distância (km)	6.45
Número de treinos dentro da semana	1.00
Frequência de prática de outros exercícios na semana	2.00
Prática de musculação	1.00
Tipo de treino	0.00
Velocidade média prevista (função objetivo)	**7.62**

Por último, e visando ainda avaliar empiricamente os benefícios da inclusão de uma etapa de otimização nessa modelagem, os resultados apresentados na Tabela V.5 evidenciam as vantagens dessa incorporação sobre as previsões do modelo. Em todos os cenários analisados, os valores ótimos de velocidade prevista superam aqueles obtidos no cenário sem otimização, indicando que o ajuste das variáveis de decisão contribui sistematicamente para a melhoria do desempenho previsto. Observa-se, contudo, que a magnitude desse ganho varia entre os modelos, sendo particularmente expressiva nos modelos segmentados por *cluster*, sobretudo no *cluster* 0, no qual a diferença entre o valor ótimo e o *baseline* é substancial. Esse resultado sugere que a abordagem segmentada é

capaz de capturar heterogeneidades relevantes entre os perfis de atletas, ampliando o potencial de melhoria por meio da otimização. Em contrapartida, no *cluster* 2, o ganho marginal indica menor sensibilidade às variáveis otimizadas, reforçando a ideia de que a efetividade da otimização depende do perfil do grupo analisado. De forma geral, esses achados demonstram que a combinação entre modelagem preditiva e otimização fornece recomendações mais eficientes do que a utilização isolada do modelo preditivo, especialmente quando aplicada a grupos mais homogêneos de atletas.

Tabela V.5: Comparação entre previsões otimizadas e *baseline*

Modelo	**Ótimo**	***Baseline***
Modelo Geral	10.64	10.27
Modelo *cluster* 0	12.11	9.07
Modelo *cluster* 1	11.65	10.28
Modelo *cluster* 2	7.62	7.53

Em síntese, os resultados da otimização reforçam a importância da personalização das recomendações de treino de acordo com o perfil do atleta, evidenciando também a relevância da escolha do modelo preditivo subjacente. Embora o modelo de Regressão Linear tenha permitido uma interpretação direta das relações entre variáveis, observou-se que, mesmo com a imposição de restrições explícitas, algumas soluções ótimas resultantes não respeitavam plenamente os limites esperados do domínio do problema. Em contraste, o modelo baseado em XGBoost produziu recomendações mais coerentes do ponto de vista fisiológico e prático, refletindo de forma mais fiel as não linearidades e interações entre as variáveis de treino. Nesse sentido, a aplicação da função `minimize` sobre um modelo mais flexível mostrou-se mais eficaz na identificação de combinações viáveis e realistas de variáveis controláveis.

Do ponto de vista prático, caso o usuário dispusesse dessas informações em um dispositivo capaz de receber os valores das variáveis controláveis e retornar a previsão de velocidade estimada, o modelo poderia ser utilizado como uma ferramenta robusta de apoio ao planejamento individual. Mesmo sob a presença de restrições operacionais, o atleta conseguiria ajustar escolhas como intensidade, distância e frequência semanal de forma orientada por previsões mais consistentes com seu perfil e histórico de treino. Assim, o sistema permitiria uma organização mais precisa e autônoma do processo de treino, conciliando desempenho, segurança e sustentabilidade ao longo do tempo.

Capítulo VI Considerações Finais

Este capítulo reúne as considerações finais do estudo, discorrendo sobre as principais limitações identificadas ao longo do desenvolvimento da pesquisa, bem como sobre os possíveis aprimoramentos metodológicos que podem fortalecer futuros desdobramentos do trabalho. Também são apresentadas direções promissoras para investigações subsequentes, contemplando oportunidades de expansão do conjunto de variáveis, aperfeiçoamento dos modelos e aprofundamento das análises. Além disso, discutem-se potenciais aplicações práticas da abordagem proposta, especialmente no suporte ao planejamento de treinos personalizados e à tomada de decisão de corredores amadores e experientes. Por fim, são sintetizadas as conclusões centrais alcançadas, destacando as contribuições teóricas e aplicadas desta investigação.

A Seção VI.1 detalha os principais fatores que podem ter influenciado os resultados obtidos, discutindo restrições relacionadas aos dados, às escolhas metodológicas e ao escopo da investigação. Na sequência, a Seção VI.2 apresenta possibilidades de aprimoramento e expansão do trabalho, com ênfase em estratégias de implementação prática, incorporação de novas variáveis e refinamento dos modelos propostos. Por fim, a Seção VI.3 sintetiza os achados mais relevantes da pesquisa, reafirmando as contribuições alcançadas e sua pertinência para o avanço do tema investigado.

VI.1 Principais Limitações

O presente estudo apresenta como uma de suas principais limitações o volume reduzido de dados utilizados na modelagem. O número limitado de participantes e de registros de treino restringe a robustez estatística das análises e compromete a generalização dos resultados para populações mais amplas de corredores. Essa limitação afeta especialmente a capacidade dos modelos de capturar a variabilidade natural entre diferentes perfis de atletas e aumenta o risco de sobreajuste, uma vez que o modelo pode se ajustar excessivamente às particularidades da amostra estudada, reduzindo sua capacidade preditiva em novas populações. Assim, a ampliação da base de dados, tanto em quantidade quanto em diversidade de indivíduos, configura-se como uma etapa essencial para aprimorar a validade externa, a confiabilidade das conclusões obtidas e a robustez do desempenho preditivo dos modelos.

Outra limitação relevante diz respeito à ausência de informações fisiológicas nas análises. As

bases utilizadas concentraram-se em variáveis pessoais, comportamentais e de desempenho derivadas dos registros de treino, sem contemplar métricas fisiológicas como frequência cardíaca ou condição respiratória. A ausência desses dados reduz o poder explicativo dos modelos, uma vez que variáveis fisiológicas são determinantes no monitoramento da carga de treino, da fadiga e da capacidade de recuperação dos atletas.

A falta de fatores biológicos mais detalhados também impacta a compreensão sobre as respostas internas dos atletas aos estímulos de treinamento. Elementos como nível de esforço percebido, eficiência respiratória e adaptação cardiovascular poderiam fornecer uma visão mais completa do desempenho, contribuindo para a construção de modelos mais realistas e fisiologicamente fundamentados.

Do ponto de vista técnico, observou-se ainda a perda de parte dos registros de treino durante o processo de leitura e consolidação dos dados. Alguns arquivos apresentaram erros de leitura ou corrompimento, o que levou à exclusão de determinadas observações do conjunto final. Embora o impacto quantitativo dessa perda tenha sido minimizado, a ausência de parte dos dados pode ter reduzido a representatividade de certos perfis de treino, influenciando os resultados e as inferências realizadas.

Outra limitação diz respeito à necessidade de inferir regras heurísticas para identificar quais registros deveriam ser classificados como atividades de corrida. Diante da ausência de um campo explícito e padronizado que discriminasse o tipo de exercício, foi necessário estabelecer critérios baseados em conhecimento prévio e padrões observados nos dados, tais como velocidade média e distância. Embora esses critérios tenham sido construídos de forma cuidadosa, eles não garantem total precisão e podem ter levado à inclusão ou exclusão equivocada de algumas atividades. Para estudos futuros, seria desejável desenvolver ou incorporar mecanismos mais robustos de classificação, como modelos específicos de detecção de modalidade esportiva ou metadados fornecidos diretamente pelos dispositivos de monitoramento, a fim de assegurar que apenas treinos de corrida sejam considerados nas análises.

Por fim, ressalta-se que nem todas as variáveis disponíveis foram exploradas. Informações como preferência de terreno para treino, histórico em outros esportes, nível de experiência competitiva, elevação, informação de bússola não foram incluídas nas análises, apesar de seu potencial contributivo para a caracterização dos perfis e para a explicação do desempenho. A inclusão dessas variáveis em estudos futuros poderá enriquecer as análises e favorecer a criação de modelos de previsão e recomendação mais completos e personalizados.

VI.2 Evoluções e Trabalhos Futuros

Como continuidade deste estudo, uma etapa natural consiste na implementação prática do modelo desenvolvido de clusterização e previsão de desempenho em um ambiente aplicado. Com o modelo já treinado e validado, propõe-se a criação de um fluxo operacional integrado em um dispositivo, como um aplicativo ou plataforma web, que permita hospedar e operacionalizar os algoritmos desenvolvidos.

Nesse cenário, quando novos usuários realizassem o preenchimento de um formulário de inscrição com suas informações pessoais e de treino, o modelo de clusterização seria automaticamente acionado para inferir o grupo (*cluster*) ao qual o atleta pertence, com base em seus atributos iniciais. Essa etapa permitiria a personalização das análises e previsões desde o primeiro contato do usuário com o sistema.

A partir desse momento, e após um período inicial de coleta de dados reais de treino (por exemplo, uma semana de atividades registradas no dispositivo), o modelo preditivo de desempenho específico para o *cluster* identificado seria utilizado para estimar a velocidade média prevista do atleta, considerando as variáveis de entrada disponíveis: tanto aquelas obtidas no questionário inicial quanto as derivadas dos registros de atividades.

Com base nessas previsões, o algoritmo de otimização poderia ser acionado para identificar as combinações ótimas das variáveis controláveis e recomendáveis (como frequência de treino, tipo de sessão, intensidade, entre outras), fornecendo assim recomendações personalizadas que maximizem o desempenho do atleta. Todo esse fluxo, que articula as etapas de coleta de dados, inferência, previsão e otimização, está representado graficamente na Figura VI.1.

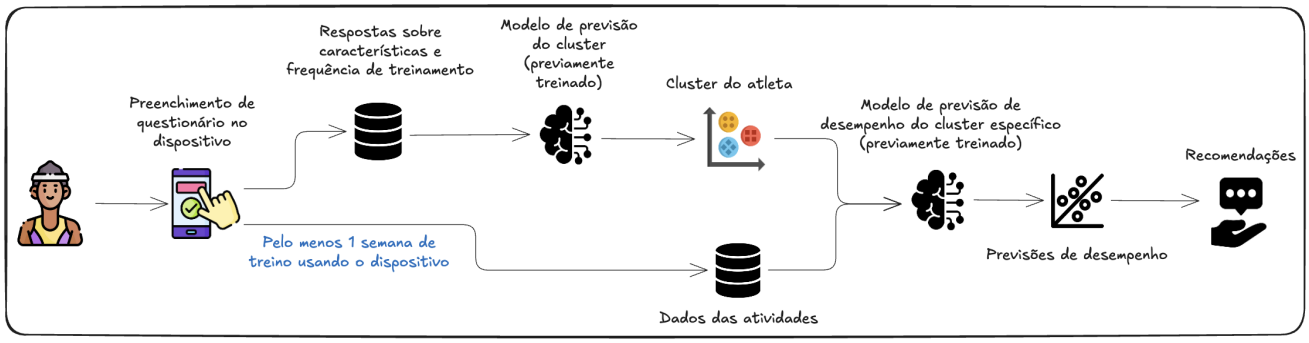


Figura VI.1: Fluxo de análise de desempenho e recomendação

Além da implementação prática, há amplo espaço para aprimoramentos metodológicos. No âmbito da previsão, seria pertinente avaliar modelos mais sofisticados, enquanto o processo de otimização e recomendação pode ser fortalecido com a incorporação de métodos heurísticos ou metaheurísticos, como algoritmos genéticos e de enxame de partículas. Tais abordagens têm potencial para lidar de forma mais eficiente com problemas que envolvem múltiplos objetivos e restrições

complexas, contribuindo para maior robustez das soluções propostas.

Estudos futuros também poderiam expandir o conjunto de variáveis preditoras, incluindo indicadores fisiológicos (como frequência cardíaca, consumo de oxigênio e lactato) e características individuais (como idade biológica e composição corporal), ampliando o poder explicativo e a capacidade de generalização do modelo. Essas extensões poderiam contribuir para a construção de um sistema inteligente de apoio ao treinamento esportivo, capaz de fornecer recomendações cada vez mais precisas e personalizadas aos diferentes perfis de corredores. Além disso, variáveis do questionário que não foram incorporadas às análises também representam oportunidades para investigações futuras. A exploração mais aprofundada desses fatores pode revelar padrões adicionais e melhorar a compreensão sobre elementos que influenciam o desempenho e a resposta ao treinamento.

Em resumo, visando aprimorar a aderência das recomendações a cenários reais, estudos futuros podem avançar no aprimoramento da modelagem preditiva, testando e incorporando variáveis adicionais que não foram consideradas na presente análise. Adicionalmente, a comparação com diferentes abordagens de modelagem, por meio da aplicação de algoritmos alternativos e arquiteturas mais sofisticadas, poderiam permitir a identificação de estruturas mais adequadas para capturar as relações subjacentes aos dados. Esses avanços tendem a reduzir limitações do modelo atual, aumentar sua robustez e ampliar sua capacidade de generalização, contribuindo para recomendações mais consistentes e alinhadas ao desempenho real dos corredores.

VI.3 Conclusões

O crescente interesse pela corrida de rua e o aumento do número de praticantes em diferentes níveis de experiência têm ampliado a necessidade de soluções personalizadas que orientem o desenvolvimento esportivo de forma eficiente e sustentável. Nesse contexto, a presente pesquisa buscou contribuir com a criação de um modelo capaz de compreender e prever o desempenho de corredores, considerando suas características individuais, hábitos e histórico de treinamento. Essa abordagem reforça a importância de estratégias baseadas em dados para apoiar a tomada de decisão dos atletas, oferecendo recomendações específicas com potencial para otimizar o desempenho e favorecer a adesão e a continuidade na prática esportiva.

Durante o desenvolvimento desta dissertação, dois estudos derivados foram produzidos e publicados no Simpósio Brasileiro de Banco de Dados (SBBD), conforme previamente apresentado na seção de Metodologia. O primeiro, intitulado “Diferenciando Perfis de Corredores por Meio de Pontos de Mudança nos Treinos” (SBBD 2024), investigou como a experiência do atleta se relaciona à estabilidade das séries temporais de velocidade, indicando menor ocorrência de pontos de mudança em corredores mais experientes. O segundo, *“Drift Detection for Identifying Training Patterns Prior to Performance Improvement in Runners”* (SBBD 2025), analisou padrões de trei-

namento que antecedem melhorias de desempenho e foi indicado ao prêmio de *best paper*. Ambos os estudos forneceram evidências empíricas importantes que orientaram a seleção das variáveis e o delineamento metodológico adotado nesta pesquisa.

O principal objetivo estudo apresentado foi desenvolver uma modelagem preditiva voltada à estimativa do desempenho de corredores e, a partir dessas previsões, aplicar algoritmos de otimização que permitissem identificar as variáveis com maior potencial de aprimoramento, viabilizando recomendações personalizadas. Para isso, conduziu-se uma pesquisa empírica com 26 corredores de diferentes níveis de maturidade esportiva, combinando dados coletados por meio de um questionário (com informações pessoais, ocupacionais e de hábitos de treino) e registros de atividades exportados de um aplicativo de corrida.

A metodologia proposta foi estruturada em três fases principais. A primeira delas envolveu a análise de clusterização, com o objetivo de identificar grupos de atletas com perfis semelhantes. Utilizando algoritmos clássicos como *K-means, Agglomerative Clustering* e DBSCAN, aplicados sobre componentes principais que explicaram 80% da variabilidade dos dados, o modelo final selecionado foi o *Agglomerative Clustering*, por apresentar o melhor desempenho no índice de silhueta. Essa abordagem possibilitou a identificação de três grupos distintos de corredores, denominados "Experientes Jovens", "Pouco Experientes" e "Experientes Veteranos", cada um com características específicas relacionadas à idade, trajetória esportiva e histórico de treino.

A seleção das variáveis de entrada utilizadas na modelagem preditiva foi conduzida de forma criteriosa, fundamentada tanto nos estudos preliminares realizados no âmbito deste próprio projeto quanto na literatura científica especializada sobre desempenho em corrida. Esse processo permitiu identificar fatores com reconhecida influência sobre a performance, envolvendo características do atleta, indicadores derivados do treinamento e elementos contextuais. Assim, buscou-se garantir que o conjunto de variáveis contemplasse dimensões relevantes e teoricamente justificadas, alinhando a modelagem desenvolvida às principais evidências empíricas já consolidadas em pesquisas anteriores.

Na segunda fase, foi desenvolvido um modelo preditivo baseado em *gradient boosting*, o XGBoost, tanto em uma configuração geral quanto em versões específicas para cada *cluster* identificado. Adicionalmente, modelos de Regressão Linear foram testados como abordagem de referência; contudo, o XGBoost apresentou desempenho superior nas métricas de avaliação consideradas, sugerindo maior capacidade de capturar relações não lineares e interações complexas entre as variáveis. A análise segmentada, por sua vez, revelou padrões diferenciados entre os *cluster*, reforçando a importância de modelos específicos para cada perfil de corredor.

Uma análise de interpretabilidade baseada nos SHAP *values* evidenciou que, de forma geral, o modelo atribui maior importância a variáveis diretamente associadas ao desempenho do treino, como velocidades mínima e máxima, distância percorrida e variabilidade do ritmo, além de fatores

históricos e contextuais, como velocidades de sessões anteriores e temperatura. No nível dos grupos, observam-se padrões distintos: no *cluster* "Experientes Jovens", predominam variáveis ligadas à intensidade do treino, indicando maior sensibilidade a picos e níveis sustentados de esforço; entre os "Pouco Experientes", destaca-se a variabilidade do ritmo e a influência do volume e do histórico recente, refletindo a instabilidade característica desse perfil; já nos "Experientes Veteranos", o histórico de desempenho e a consistência do treinamento assumem papel central, com maior relevância para métricas de regularidade, volume e recuperação. Complementarmente, uma análise de sensibilidade também foi conduzida, apresentando resultados em consonância com os padrões identificados na interpretabilidade por SHAP, o que reforça a consistência das relações observadas. Em conjunto, esses resultados reforçam a adequação da segmentação proposta e indicam que o modelo sugere diferenças estruturais na resposta ao treino entre perfis distintos de corredores.

Na terceira etapa, foi implementado um algoritmo de otimização com o objetivo de identificar combinações de variáveis controláveis que maximizassem a velocidade média prevista como *proxy* de desempenho, respeitando as restrições impostas pelo modelo. No cenário geral, a otimização resultou em uma velocidade média prevista de 10,64 km/h, associada a um perfil de treino moderado, caracterizado por volume e frequência sustentáveis, prática de musculação e ausência de treinos intervalados, sugerindo que, ao considerar os atletas de forma conjunta, o modelo privilegia estratégias conservadoras e consistentes. A análise segmentada por *clusters* revelou padrões distintos entre os grupos: os "Experientes Jovens" apresentaram o maior valor ótimo de velocidade média (12,11 km/h), associado a maior diversidade de estímulos e intensidade moderada; os "Pouco Experientes" atingiram um desempenho intermediário (11,65 km/h), sugerindo que a progressão gradual de volume e regularidade é central nesse perfil; e os "Experientes Veteranos" exibiram a menor velocidade média prevista (7,62 km/h), coerente com estratégias focadas em controle de carga, sustentabilidade e prevenção de sobrecarga. Em conjunto, esses resultados apontam que a segmentação permite capturar estratégias ótimas de treino específicas a cada perfil, que não seriam observáveis em um modelo agregado.

Adicionalmente, os resultados da otimização foram comparados a um cenário *baseline*, sem ajuste das variáveis de decisão, evidenciando ganhos consistentes de desempenho em todos os modelos analisados. Observou-se que a incorporação da etapa de otimização esteve associada a velocidades médias previstas superiores, com ganhos mais expressivos nos modelos segmentados por *cluster*, em especial no grupo de "Experientes Jovens", enquanto o impacto é mais moderado entre os "Experientes Veteranos". Esses achados reforçam que a efetividade da otimização depende do perfil do atleta e da homogeneidade do grupo analisado. Como forma de fortalecer a robustez das conclusões, também foram conduzidas análises de sensibilidade na etapa de otimização, corroborando a estabilidade dos resultados e a consistência dos padrões observados.

De modo geral, os resultados obtidos sugerem a viabilidade do uso de modelos de aprendizado de máquina para compreender padrões de desempenho e apoiar a prescrição personalizada de treino. A integração entre técnicas de clusterização, regressão e otimização mostrou-se uma abordagem promissora para sistemas inteligentes de recomendação no contexto esportivo, especialmente por permitir a identificação de perfis de atletas e a sugestão de ajustes individualizados com base em dados objetivos.

Entretanto, reconhece-se que o estudo ainda possui limitações, especialmente no que diz respeito ao tamanho da amostra e à ausência de variáveis fisiológicas, que poderiam enriquecer as análises. Mesmo assim, os resultados oferecem evidências iniciais relevantes sobre o potencial da modelagem preditiva aplicada ao esporte, estabelecendo uma base sólida para o desenvolvimento de sistemas mais robustos e integrados.

Como encaminhamento futuro, propõe-se a implementação prática do modelo em um ambiente digital, como um aplicativo ou plataforma *online*, capaz de hospedar os algoritmos e operacionalizar o fluxo completo de inferência, previsão e otimização. Tal sistema permitiria que os usuários inserissem suas informações, fossem automaticamente classificados em *clusters* e recebessem recomendações personalizadas de treino com base em suas características e histórico de desempenho. Além disso, a expansão do modelo para incluir variáveis fisiológicas, psicológicas e contextuais poderá consolidar uma ferramenta inteligente de apoio ao treinamento, fortalecendo a convergência entre ciência de dados e performance esportiva.

Em síntese, este trabalho evidencia que a aplicação de técnicas de aprendizado de máquina e otimização pode desempenhar um papel estratégico na personalização do treinamento de corredores, auxiliando atletas de diferentes níveis a compreender melhor seu desempenho e a tomar decisões mais assertivas. Ao combinar análise de dados, modelagem preditiva e otimização, abre-se caminho para o desenvolvimento de soluções inovadoras que aproximam a ciência da prática esportiva, contribuindo para o avanço significativo na forma como se interpreta, planeja e melhora o desempenho humano no esporte.

Referências

Berndsen, J., Lawlor, A., and Smyth, B. Running with recommendation. In *CEUR Workshop Proceedings*, volume 1953, pages 18 – 21, 2017.

Berndsen, J., Smyth, B., and Lawlor, A. Pace my race: recommendations for marathon running. In *Proceedings of the 13th ACM Conference on Recommender Systems*, page 246–250, NY, USA. Association for Computing Machinery, 2019.

Besomi, M., Leppe, J., Di Silvestre, M. C., and Setchell, J. Serun study: Development of running profiles using a mixed methods analysis. *PLOS ONE*, 13(7):1–18, 2018.

Billat, V., DEMARLE, A., Slawinski, J., Paiva, M., and Koralsztein, J. P. Physical and training characteristics of top-class marathon runners. *Medicine and science in sports and exercise*, 33:2089–97, 2001.

Bontempi, G., Ben Taieb, S., and Le Borgne, Y.-A. *Machine Learning Strategies for Time Series Forecasting*, volume 138, 2013.

Bushra, A. A. and Yi, G. Comparative analysis review of pioneering dbscan and successive density-based clustering algorithms. *IEEE Access*, 9:87918–87935, 2021.

Casado, A., Hanley, B., Jiménez-Reyes, P., and Renfree, A. Pacing profiles and tactical behaviors of elite runners. *Journal of Sport and Health Science*, 10(5):537 – 549, 2021.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794. ACM, 2016.

Cohen, M. C., Leung, N.-H. Z., Panchamgam, K., Perakis, G., and Smith, A. The impact of linear optimization on promotion planning. *Operations Research*, 65(2):446–468, 2017.

Corbett, D., Sweeting, A., and Robertson, S. A change point approach to analysing the match activity profiles of team-sport athletes. *Journal of Sports Sciences*, 37, 2019.

El Helou, N., Tafflet, M., Berthelot, G., Tolaini, J., Marc, A., Guillaume, M., Hausswirth, C., and Toussaint, J.-F. Impact of environmental parameters on Marathon running performance. *PLOS ONE*, 7(5), 2012.

El-Kassabi, H. T., Khalil, K., and Serhani, M. A. Deep learning approach for forecasting athletes' performance in sports tournaments. In *ACM International Conference Proceeding Series*, pages 203 – 208, 2020.

Ely, M. R., Martin, D. E., Cheuvront, S. N., and Montain, S. J. Effect of ambient temperature on marathon pacing is dependent on runner ability. *Medicine and Science in Sports and Exercise*, 40(9):1675 – 1680, 2008.

Gómez-Molina, J., Ogueta-Alday, A., Camara, J., Stickley, C., Rodríguez-Marroyo, J. A., and García-López, J. Predictive variables of half-marathon performance for male runners. *Journal of Sports Science and Medicine*, 16(2):187 – 194, 2017.

Gosztyla, A., Edwards, D., Quinn, T., and Kenefick, R. The impact of different pacing strategies on five-kilometer running time trial performance. *Journal of strength and conditioning research / National Strength & Conditioning Association*, 20:882–6, 2006.

Hespanhol Junior, L., Pillay, J., and van Mechelen, W. Meta-analyses of the effects of habitual running on indices of health in physically inactive adults. *Sports Medicine*, 45:1455–1468, 2015.

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., and Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

Jauhiainen, S., Pohl, A., Äyrämö, S., Kauppi, J.-P., and Ferber, R. A hierarchical cluster analysis to determine whether injured runners exhibit similar kinematic gait patterns. *Scandinavian Journal of Medicine & Science in Sports*, 30, 2020.

Johansson, M., Atterfors, J., and Lamm, J. Pacing patterns of half-marathon runners: An analysis of ten years of results from gothenburg half marathon. *International Journal of Computer Science in Sport*, 22:124–138, 2024.

Kakaew, S. and U-tapao, C. An optimization model for the maximized profit of housing development project: A case study of bangkok and perimeter, thailand. *Applied Sciences*, 13(9), 2023.

Knechtle, B., Knechtle, P., Barandun, U., Rosemann, T., and Lepers, R. Predictor variables for half marathon race time in recreational female runners. *Clinics*, 66(2):287 – 291, 2011.

Koegel, J., Huerta, S., Gambietz, M., Ullrich, M., Heyde, C., Dorschky, E., and Eskofier, B. Clustering runners, response to different midsole stack heights: a field study. *sensors*, 24(14), 2024.

Komitova, R., Raabe, D., Rein, R., and Memmert, D. Time Series Data Mining for Sport Data: a Review. *IJCSS*, 21(2):17 – 31, 2022.

Koronios, K., Psiloutsikou, M., and Kriemadis, A. Motives and constraints of participants in running events. *Education and Training*, 60(5):443 – 457, 2017.

Kraft, D. A software package for sequential quadratic programming. Technical Report DFVLR-FB 88-28, DLR German Aerospace Center, Germany, 1988.

Kumar, S. A survey on optimization algorithms: Challenges and future opportunities. *international journal of advanced research in computer science*, 15:45–51, 2024.

Lerebourg, L., Saboul, D., Clémençon, M., and Coquart, J. B. Prediction of Marathon Performance using Artificial Intelligence. *International Journal of Sports Medicine*, 44(05):352–360, 2023.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Lundy, L. and Reilly, R. B. Demographics, culture and participatory nature of multi-marathoning—an observational study highlighting issues with recommendations. *PLOS ONE*, 19(5):1–14, 2024.

Maulud, D. and Abdulazeez, A. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1:140–147, 2020.

Mišić, V. V. Optimization of tree ensembles. *Operations Research*, 68(5):1605–1624, 2020.

Muñoz, I., Moreno Pérez, D., Moreno, P., González, C., and Esteve, J. Prediction of race pace in long distance running from blood lactate concentration around race pace. *Journal of Human Sport and Exercise*, 7:763–769, 2012.

Nakayama, Y., Kudo, K., and Ohtsuki, T. Variability and fluctuation in running gait cycle of trained runners and non-runners. *Gait & posture*, 31:331–5, 2010.

Nikolaidis, P., Rosemann, T., and Knechtle, B. Sex differences in the age of peak marathon race time. *The Chinese journal of physiology*, in print, 2017.

Ogasawara, E. S., Salles, R., Porto, F., and Pacitti, E. *Event Detection in Time Series*. Synthesis Lectures on Data Management (SLDM). Springer, 2025.

Olive, D. *Linear regression*, 2017.

Omran, M., Engelbrecht, A., and Salman, A. An overview of clustering methods. *Intell. Data Anal.*, 11:583–605, 2007.

Parra-Camacho, D., González-Serrano, M. H., González-García, R. J., and Calabuig Moreno, F. Sporting habits of urban runners: Classification according to their motivation. *International Journal of Environmental Research and Public Health*, 16(24), 2019.

Pérez, I. M., Pérez, D. M., González, C. C., and Esteve-Lanao, J. Prediction of race pace in long distance running from blood lactate concentration around race pace. *Journal of Human Sport and Exercise*, 7(4):763 – 769, 2012.

Rebbouj, M. and Lotfi, S. Athlete performance prediction using intelligent reporting and regression model analysis: A generative approach for training planning. *Journal of Physical Education and Sport*, 24:1927–1932, 2024.

Reusser, M., Sousa, C., Villiger, E., Alvero Cruz, J. R., Hill, L., Rosemann, T., Nikolaidis, P., and Knechtle, B. Increased participation and decreased performance in recreational master athletes in "berlin marathon" 1974–2019. *Frontiers in Physiology*, 12:631237, 2021.

Rodriguez Rivadulla, A., Chen, X., Cazzola, D., Trewartha, G., and Preatoni, E. Clustering analysis across different speeds reveals two distinct running techniques with no differences in running economy. *Sports biomechanics*, pages 1–24, 2024.

Shu, D., Wang, J., Zhou, T., Chen, F., Meng, F., Wu, X., Zhao, Z., and Dai, S. Prediction of half-marathon performance of male recreational marathon runners using nomogram. *BMC Sports Science, Medicine and Rehabilitation*, 16, 2024.

Smyth, B. Recommender systems: A healthy obsession. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9790–9794, 2019.

Smyth, B., Lawlor, A., Berndsen, J., and Feely, C. Recommendations for marathon runners: on the application of recommender systems and machine learning to support recreational marathon runners. *User Modeling and User-Adapted Interaction*, 32(5):787 – 838, 2022.

Suwankan, S., Suwankan, S., Theanthong, A., and Kemarat, S. Prediction of 10 km running time by physical and training characteristics in recreational runners, 2024.

Tanda, G. Prediction of marathon performance time on the basis of training indices. *Journal of Human Sport and Exercise*, 6(3):511 – 520, 2011.

Teune, B., Woods, C., Sweeting, A., Inness, M., and Robertson, S. A method to inform team sport training activity duration with change point analysis. *PLOS ONE*, 17(3):1–11, 2022.

Thuany, M., Gomes, T. N., and Almeida, M. Is there any difference between "amateur"and "recreational"runners? a latent class analysis. *Motriz. Revista de Educação Física*, 2020.

Thuany, M., Gomes, T. N., ESTEVAM, L., and Almeida, M. *Crescimento no número de corridas de rua e perfil dos participantes no Brasil*, 2021a.

Thuany, M., Knechtle, B., Rosemann, T., Almeida, M., and Gomes, T. N. Running around the country: An analysis of the running phenomenon among brazilian runners. *International Journal of Environmental Research and Public Health*, 18:6610, 2021b.

Thuany, M., Vieira, D., Villiger, E., Gomes, T. N., Weiss, K., Nikolaidis, P. T., Sousa, C. V., Scheer, V., and Knechtle, B. An analysis of the São Silvestre race between 2007–2021: An increase in participation but a decrease in performance. *Sports Medicine and Health Science*, 5(4):277 – 282, 2023.

Tito, N., Paixão, B., Tavares, L., Ogasawara, E., and Amorim, G. Diferenciando perfis de corredores por meio de pontos de mudança nos treinos. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 834–840, Porto Alegre, Brasil. SBC, 2024.

Tito, N., Tavares, L., and Amorim, G. Drift detection for identifying training patterns prior to performance improvement in runners. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 795–801, Porto Alegre, RS, Brasil. SBC, 2025.

Tokuda, E. K., Comin, C. H., and da F. Costa, L. Revisiting agglomerative clustering. *Physica A: Statistical Mechanics and its Applications*, 585:126433, 2022.

V., D., Anguraj, D. K., and Mantha, S. An efficient recommendation system for athletic performance optimization by enriched grey wolf optimization. *Personal and Ubiquitous Computing*, 27:1–12, 2022.

Van den Berghe, P., Gosseries, M., Gerlo, J., Lenoir, M., Leman, M., and De Clercq, D. Change-point detection of peak tibial acceleration in overground running retraining. *Sensors*, 20(6), 2020.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt,

P., and SciPy 1. 0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Medicine*, 17:261–272, 2020.

Wang, S., Scheider, S., Sporrel, K., Deutekom, M., Timmer, J., and Kröse, B. What are good situations for running? a machine learning study using mobile and geographical data. *Frontiers in Public Health*, 8, 2021.

Wiecha, S., Kasiak, P. S., Cieśliński, I., Maciejczyk, M., Mamcarz, A., and Śliż, D. Modeling physiological predictors of running velocity for endurance athletes. *Journal of Clinical Medicine*, 11(22), 2022.

Yong, W., Lingyun, P., and Jia, W. Statistical analysis and ARMA modeling for the big data of marathon score. *Science and Sports*, 35(6):375 – 385, 2020.

Zach, S., Xia, Y., Zeev, A., Arnon, M., Choresh, N., and Tenenbaum, G. Motivation dimensions for running a marathon: A new model emerging from the motivation of marathon scale (moms). *Journal of Sport and Health Science*, 6(3):302–310, 2017.