Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Desenvolvimento de um Sistema Embarcado para Monitoramento e Apoio a Corredores Amadores com Técnicas de Aprendizagem de Máquina

Núbia Ribeiro Naliatti de Mello

JUIZ DE FORA
JANEIRO, 2026

Desenvolvimento de um Sistema Embarcado para Monitoramento e Apoio a Corredores Amadores com Técnicas de Aprendizagem de Máquina

Núbia Ribeiro Naliatti de Mello

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Luciana Conceição Dias Campos

Coorientador: Victor Ströele de Andrade Menezes

JUIZ DE FORA

JANEIRO, 2026

Desenvolvimento de um Sistema Embarcado para Monitoramento e Apoio a Corredores Amadores com Técnicas de Aprendizagem de Máquina

Núbia Ribeiro Naliatti de Mello

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Luciana Conceição Dias Campos
Doutora em Engenharia Elétrica (PUC-Rio)

Victor Ströele de Andrade Menezes
Doutor em em Engenharia de Sistemas e Computação (UFRJ)

André Luiz de Oliveira
Doutor em Ciências da Computação e Matemática Computacional (ICMC/USP)

Marcelo Caniato Renhe
Doutor em Computação Gráfica (UFRJ)

JUIZ DE FORA

20 DE JANEIRO, 2026

Aos meus amigos e irmãos.

Aos pais, pelo apoio e sustento.

Resumo

A prática da corrida tem se popularizado entre pessoas que buscam qualidade de vida e saúde, especialmente entre corredores amadores. No entanto, a ausência de acompanhamento profissional pode aumentar o risco de lesões e dificultar a evolução. Este trabalho propõe o desenvolvimento de um sistema de apoio inteligente voltado a corredores amadores, utilizando técnicas de aprendizado de máquina aplicadas à análise de dados fisiológicos e de desempenho. O projeto constitui uma continuação do trabalho desenvolvido por Pedro Henrique Almeida Cardoso Reis em sua monografia "Uso de Machine Learning no Esporte: Apoio Inteligente para Corredores não Profissionais" (REIS, 2025), aproveitando e estendendo parte de sua infraestrutura, incluindo uma das interfaces desenvolvidas anteriormente. Neste trabalho, foram analisados dados coletados de dois voluntários que utilizam dispositivos vestíveis comerciais, como *smartwatches*, ou aplicativos de corrida, com o objetivo de ajustar os modelos preditivos utilizados no trabalho precedente, oferecendo assim, percepções relevantes personalizadas ao corredor, auxiliando na otimização do treino e na prevenção de sobrecargas. Os resultados demonstraram o potencial do aprendizado de máquina para a análise de dados de corrida, destacando o algoritmo de *Gradient Boosting* como o de melhor desempenho entre os modelos avaliados. Observou-se uma pequena diferença de desempenho em relação aos resultados obtidos no estudo precedente, mantendo-se, entretanto, um padrão semelhante, no qual modelos de *ensemble* apresentaram desempenho superior, enquanto algoritmos como *Support Vector Regression* (SVR) e *K-Nearest Neighbors* (KNN) apresentaram desempenho inferior. Esses resultados reforçam a viabilidade do uso de técnicas de aprendizado de máquina como ferramentas acessíveis e eficazes para o apoio a corredores não profissionais.

Palavras-chave: aprendizado de máquina, corredores amadores, *e-health*, análise de dados, desempenho esportivo.

Abstract

Running has become increasingly popular among people seeking quality of life and health, especially among amateur runners. However, the lack of professional guidance can increase the risk of injuries and hinder performance improvement. This work proposes the development of an intelligent support system aimed at amateur runners, using machine learning techniques applied to the analysis of physiological and performance data. The project is a continuation of the work developed by Pedro Henrique Almeida Cardoso Reis in his monograph "Use of Machine Learning in Sports: Intelligent Support for Non-Professional Runners" (REIS, 2025), reusing and extending part of its infrastructure, including one of the previously developed interfaces. In this study, data collected from two volunteers who use commercial wearable devices, such as smartwatches, and running applications were analyzed in order to build predictive models capable of providing personalized and relevant insights to runners, supporting training optimization and injury prevention. The results demonstrated the potential of machine learning for the analysis of running data, with the Gradient Boosting algorithm achieving the best performance among the evaluated models. A small difference in performance was observed in comparison with the results of the previous study, while maintaining a similar pattern in which ensemble models outperformed others, whereas algorithms such as Support Vector Regression (SVR) and K-Nearest Neighbors (KNN) showed inferior performance. These results reinforce the feasibility of using machine learning techniques as accessible and effective tools to support non-professional runners.

Keywords: machine learning; amateur runners; e-health; data analysis; sports performance..

Agradecimentos

Gostaria de agradecer aos meus pais, Rildo e Viviane, e a minha irmã, Sofia, pelo apoio, encorajamento, conselhos e sustento durante esses longos anos.

A professora Luciana e ao professor Victor, pela orientação, amizade e principalmente, pela paciência, sem a qual este trabalho não se realizaria, destacando-se ainda a colaboração do professor Victor como um dos voluntários na disponibilização dos dados utilizados neste estudo.

Em seguida, gostaria de estender minha gratidão à minha psicóloga, Ludmila Rachid, que foi mais do que essencial para conclusão dessa etapa. Sem sua presença, nada disso seria possível.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o nosso enriquecimento pessoal e profissional.

Agradeço, por fim, Rodolfo Granato que também forneceu gentilmente seus dados de corrida, permitindo a realização desta pesquisa.

"Lembra que o sono é sagrado e alimenta de horizontes o tempo acordado de viver".

Beto Guedes (Amor de Índio)

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

DCC	Departamento de Ciência da Computução
UFJF	Universidade Federal de Juiz de Fora
ML	Machine Learning
MAE	Mean Absolut Error
R^2	Coeficiente de Determinação
RMSE	Root Mean Squared Error
MAPE	Mean Absolute Percentage Error

1 Introdução

1.1 Apresentação do tema

Este trabalho dá continuidade à pesquisa iniciada por Pedro Henrique Almeida Cardoso Reis em sua monografia "Uso de Machine Learning no Esporte: Apoio Inteligente para Corredores não Profissionais" (REIS, 2025), que teve como foco principal o desenvolvimento de um sistema preditivo baseado em dados de um único corredor. A proposta atual amplia essa abordagem ao incorporar novos dados, promovendo uma base de dados mais diversa e representativa. Com isso, busca-se aumentar a robustez dos modelos preditivos desenvolvidos, permitindo maior generalização e adequação a contextos práticos do cotidiano esportivo. O uso de tecnologias inteligentes no contexto esportivo vem crescendo rapidamente, sobretudo com o avanço de dispositivos vestíveis e aplicativos. Corredores amadores, cada vez mais engajados em melhorar seu desempenho e bem-estar, representam um público que pode se beneficiar fortemente de ferramentas baseadas em análise de dados. Este projeto propõe o desenvolvimento de um sistema de apoio inteligente que utiliza dados coletados de forma prática, com foco em fornecer previsões e orientações personalizadas com base no histórico de desempenho dos usuários.

Segundo relatório da plataforma Strava, a corrida foi o esporte mais praticado no mundo em 2024, e o Brasil aparece como o segundo país com o maior número de corredores, com mais de 19 milhões de praticantes. Além disso, houve um crescimento de 29% no número de corridas de rua oficiais realizadas no país entre 2023 e 2024, totalizando 2.827 eventos (STRAVA, 2024). Esse aumento não se restringe a atletas profissionais. A corrida de rua tornou-se um fenômeno cultural e acessível, atraindo pessoas em busca de saúde, lazer e qualidade de vida.

Com o avanço de dispositivos vestíveis (*smartwatches*, sensores, aplicativos), corredores amadores passaram a ter acesso a métricas como frequência cardíaca, velocidade, distância percorrida, temperatura corporal e gasto calórico. No entanto, a interpretação desses dados ainda é limitada para quem não dispõe de acompanhamento técnico. Sis-

temas embarcados com técnicas de aprendizagem de máquina podem transformar esses dados em informações úteis, oferecendo *insights* personalizados sobre desempenho, prevenção de lesões e padrões de comportamento durante treinos e provas.

Apesar da ampla disponibilidade de dados, corredores amadores enfrentam dificuldades para compreender a relevância e a funcionalidade das variáveis coletadas. Diferentemente de atletas profissionais, corredores amadores raramente contam com treinadores ou equipes médicas para interpretar métricas e ajustar treinos. Há potencial para identificar padrões de comportamento e desempenho (como variações de ritmo, sinais de fadiga ou risco de lesão), mas isso exige ferramentas inteligentes capazes de analisar grandes volumes de dados (OLIVEIRA; SANTOS, 2022). Assim, o problema central é a lacuna entre a coleta de dados e sua utilização efetiva para apoiar corredores amadores. O desenvolvimento de um sistema embarcado com técnicas de aprendizagem de máquina busca preencher essa lacuna, oferecendo monitoramento inteligente e recomendações personalizadas que transformem dados brutos em informações acionáveis para melhorar o desempenho, a segurança e a motivação.

1.2 Justificativa/Motivação

A corrida é uma prática física acessível, que pode ser realizada gratuitamente em espaços públicos como ruas e parques, o que contribui para sua crescente popularidade entre pessoas que buscam melhorar a saúde e a qualidade de vida. Além dos benefícios fisiológicos e psicológicos já amplamente reconhecidos, sua simplicidade e baixo custo inicial favorecem a adesão de praticantes iniciantes. No entanto, o acompanhamento profissional — importante para prevenir lesões e orientar o progresso adequado — normalmente envolve custos financeiros que nem todos os corredores amadores estão dispostos ou aptos a assumir.

Com o avanço dos dispositivos vestíveis e dos aplicativos de monitoramento, tornou-se possível coletar uma ampla variedade de dados relevantes durante os treinos, como frequência cardíaca, ritmo, distância e tempo. No entanto, esses dados, por si só, têm utilidade limitada sem uma análise contextualizada. Dessa forma, propõe-se, neste trabalho, o desenvolvimento de um sistema que utilize técnicas de aprendizado de máquina para interpretar esses dados de forma inteligente, auxiliando o corredor em sua evolução

e na prevenção de lesões. O sistema busca ser acessível e de fácil utilização, promovendo a democratização do uso da tecnologia no treinamento esportivo, especialmente entre praticantes não profissionais.

Este trabalho se justifica pela necessidade de um sistema que utilize aprendizado de máquina para analisar esses dados e fornecer orientações personalizadas aos corredores amadores, contribuindo para a prevenção de lesões e para a otimização dos treinos. Assim, a tecnologia pode tornar o esporte mais seguro e eficiente para quem não dispõe de suporte profissional, contribuindo para a saúde e o bem-estar dessa população.

1.3 Objetivos

1.3.1 Objetivos Gerais

Ampliar a base de dados para ajustar modelos de aprendizado de máquina embarcados em dispositivos vestíveis ou aplicativos em dispositivos móveis para oferecer suporte à prática de corrida entre corredores amadores, visando maior eficácia, com melhor desempenho em termos de acurácia e robustez na generalização dos resultados.

1.3.2 Objetivos Específicos

- Realizar um levantamento bibliográfico sobre a aplicação do aprendizado de máquina no esporte, especialmente na corrida.
- Definir um protocolo de coleta individual de dados com voluntários.
- Realizar o pré-processamento, o tratamento e a organização dos dados.
- Aplicar e avaliar diferentes modelos de *machine learning* embarcados.
- Validar os modelos por meio de métricas estatísticas e de testes práticos.

1.4 Organização do trabalho

Este Trabalho de Conclusão de Curso está organizado da seguinte forma: O **Capítulo 1** apresenta o tema do trabalho, bem como as motivações que justificam sua realização e

os objetivos gerais e específicos propostos. No **Capítulo 2**, são apresentados os principais conceitos teóricos que sustentam o desenvolvimento do estudo. O **Capítulo 3** aborda a revisão da literatura, reunindo trabalhos e pesquisas relacionados ao tema. No **Capítulo 4** é descrito o desenvolvimento do trabalho, incluindo a metodologia adotada, as análises realizadas e os resultados obtidos. Por fim, o **Capítulo 5** apresenta as conclusões alcançadas e sugestões para trabalhos futuros.

2 Fundamentação Teórica

Nesta fundamentação teórica, são discutidos três pilares centrais para a compreensão e desenvolvimento deste estudo e analisadas suas inter-relações, buscando compreender como a combinação dessas áreas pode contribuir para a melhoria do desempenho e da segurança na corrida amadora. A Seção 2.1 aborda o conceito de *e-Health*, a Seção 2.2 discute o perfil dos corredores amadores, e a Seção 2.3 apresenta técnicas de aprendizado de máquina.

1. *e-Health* e Monitoramento Esportivo: O conceito de *e-health* (saúde digital) envolve o uso de tecnologia para promover a saúde. No contexto esportivo, dispositivos como *smartwatches* permitem o acompanhamento contínuo de variáveis fisiológicas e de desempenho.

2. Corredores Amadores e Treinamento Personalizado: Diferentemente de atletas profissionais, corredores amadores nem sempre contam com orientação especializada. Soluções tecnológicas podem suprir parte dessa lacuna, oferecendo *insights* sobre sua prática esportiva.

3. Aprendizado de Máquina (*Machine Learning*): As técnicas de aprendizado de máquina, como regressão, árvores de decisão e florestas aleatórias, são capazes de identificar padrões ocultos em grandes volumes de dados e produzir previsões úteis, como estimativa do tempo de corrida, risco de fadiga ou sugestões de ritmo.

2.1 *e-Health* e Monitoramento Esportivo

O uso de tecnologias digitais aplicadas à saúde, frequentemente referido como *e-Health*, representa uma transformação significativa na forma como os serviços de promoção, prevenção e acompanhamento de condições de saúde são prestados. Segundo Oh et al. (2005), *e-Health* engloba "o uso de tecnologias de informação e comunicação em apoio à saúde e aos sistemas de saúde". No contexto esportivo, esse conceito tem sido ampliado para

incluir dispositivos portáteis capazes de coletar dados fisiológicos e de desempenho em tempo real, como frequência cardíaca, velocidade, distância percorrida e variabilidade do ritmo.

Com a popularização de *wearables* — em especial *smartwatches* e sensores vestíveis — tornou-se possível o monitoramento contínuo de variáveis que, até pouco tempo atrás, eram restritas a laboratórios esportivos ou avaliações clínicas especializadas. Esses dispositivos não apenas registram dados, mas também fornecem feedback imediato ao usuário, abrindo espaço para que indivíduos comuns possam acompanhar sua saúde e seu desempenho de forma autônoma. Segundo Patel, Asch e Volpp (2012), o principal benefício do monitoramento digital é a capacidade de transformar sinais fisiológicos em informações acionáveis, potencializando comportamentos saudáveis e estratégias de treinamento mais eficazes.

Além disso, pesquisas mostram que a utilização de tecnologias de *e-Health* no esporte pode melhorar a adesão ao treinamento e facilitar a autoavaliação de desempenho, fatores que são particularmente relevantes para praticantes sem acompanhamento profissional constante (WANG; XU, 2020) . Isso ocorre porque a visualização e interpretação de dados em plataformas conectadas permitem ao usuário entender melhor seus padrões de resposta ao esforço e adaptar seus hábitos de forma informada.

2.2 Corredores Amadores e Treinamento Personalizado

Os corredores amadores compõem um grupo diverso de praticantes que buscam, na corrida — muitas vezes de rua — benefícios relacionados à saúde, bem-estar e realização pessoal, sem necessariamente receber remuneração ou ter vínculo profissional com o esporte (ANTONIO; LAUX, 2022). Estudos indicam que a corrida amadora vem crescendo significativamente em popularidade, com muitos indivíduos treinando de maneira autônoma e com frequência e volume variados (STUDY, 2020; NETTO, 2017). Essa heterogeneidade de perfis reflete também diferenças nas necessidades de orientação e de suporte ao treinamento, visto que muitos corredores não contam com acompanhamento profissional

constante, o que pode influenciar tanto o desempenho quanto a prevenção de lesões (AL., 2021; AL., 2025).

A literatura aponta que intervenções de treinamento personalizado, conduzidas por educadores físicos ou por meio de aplicações digitais com adaptação automática das sessões, podem melhorar os resultados dos corredores amadores ao ajustar os estímulos de treino às características individuais e ao nível de condicionamento de cada praticante (AL., 2021; AL., 2020a). Por exemplo, estudos sobre aplicativos de corrida revelam que mecanismos de personalização baseados em dados de biofeedback [1] e métricas de GPS são capazes de adaptar automaticamente os planos de treino às demandas fisiológicas do corredor, incrementando a motivação e a percepção de um treino individualizado (AL., 2020a; AL., 2020b).

Além disso, pesquisas demonstram que a prescrição adequada do treinamento, incluindo variáveis como volume, intensidade e recuperação, está associada à melhoria do desempenho de corredores recreacionais e à redução de riscos associados ao *overtraining* ou à progressão inadequada da carga de exercício (AL., 2021; STUDY, 2020).

Dessa forma, a personalização do treinamento emerge como um elemento central para atender às necessidades específicas de corredores amadores, especialmente diante da variabilidade de objetivos, níveis de experiência e respostas fisiológicas que este grupo apresenta. Ao integrar abordagens tradicionais de assessoria esportiva com recursos tecnológicos de monitoramento e de feedback, é possível oferecer um suporte mais eficaz e alinhado às metas individuais de cada corredor.

2.3 Aprendizado de Máquina (*Machine Learning*)

Machine Learning (ML) é uma subárea da inteligência artificial que permite que sistemas aprendam padrões e realizem previsões com base em dados, sem necessidade de programação explícita para cada tarefa (MITCHELL, 1997). Os algoritmos de ML podem ser divididos em três categorias principais: aprendizado supervisionado, não supervisionado e por reforço (GOODFELLOW; BENGIO; COURVILLE, 2016).

[1] Biofeedback é uma técnica que utiliza dispositivos eletrônicos para monitorar e fornecer, em tempo real, informações sobre processos fisiológicos do indivíduo — como frequência cardíaca, atividade muscular ou respiração — permitindo maior consciência corporal e autorregulação dessas funções.

No aprendizado supervisionado, modelos são treinados com dados rotulados, permitindo que façam previsões sobre novas amostras. Técnicas como regressão linear, árvores de decisão e redes neurais artificiais são exemplos comuns. Já o aprendizado não supervisionado busca estruturar ou agrupar dados não rotulados, utilizando métodos como *clustering* e redução de dimensionalidade (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O aprendizado por reforço, por sua vez, baseia-se na interação do agente com o ambiente, otimizando suas ações para maximizar uma função de recompensa (SUTTON; BARTO, 2018).

Recentemente, a aplicação de ML em saúde e esportes tem recebido destaque. Sistemas de *e-health* incorporam ML para o monitoramento contínuo de indicadores fisiológicos, a análise de desempenho e a prevenção de lesões (TOPOL, 2019). Em particular, o monitoramento esportivo de corredores amadores pode se beneficiar da análise preditiva, utilizando dados como frequência cardíaca, distância percorrida, cadência e condições ambientais para estimar tempos futuros de corrida e otimizar planos de treino (BUCHHEIT; LAURSEN, 2013).

Além disso, modelos de ML podem ser integrados ao treinamento personalizado, considerando a variabilidade individual de cada atleta. Por exemplo, algoritmos de regressão e redes neurais podem correlacionar parâmetros fisiológicos com desempenho, permitindo ajustes dinâmicos nos treinos (BACA; KORNFEIND, 2011). Essa abordagem promove uma estratégia de treinamento mais eficiente e segura, especialmente para corredores amadores que buscam melhorar seus tempos sem aumentar o risco de lesões.

2.3.1 Algoritmos de *Machine Learning* Utilizados

Nesta seção são apresentados os algoritmos de *Machine Learning* empregados neste trabalho para a tarefa de regressão, cujo objetivo é estimar o tempo necessário para percorrer uma determinada distância de corrida a partir de dados históricos. A Figura 2.1 ilustra, de forma conceitual, os diferentes mecanismos de busca adotados pelos seis algoritmos de aprendizado de máquina utilizados neste trabalho

- ***Gradient Boosting*:** Método de *ensemble* que constrói modelos de forma sequencial, em que cada novo modelo busca corrigir os erros cometidos pelos modelos

anteriores. A técnica baseia-se na minimização de uma função de perda por meio de descida do gradiente e apresenta alto desempenho em tarefas de regressão, sendo amplamente adotada em aplicações práticas (FRIEDMAN, 2001).

- ***Linear Regression*:** A Regressão Linear é um dos modelos mais simples e amplamente utilizados em problemas de regressão. Seu objetivo é modelar a relação entre variáveis independentes e uma variável dependente por meio de uma combinação linear dos atributos de entrada. Apesar de sua simplicidade, esse modelo serve como uma importante linha de base (baseline) para comparação com modelos mais complexos (MONTGOMERY; PECK; VINING, 2012).

- ***Decision Tree*:** As Árvores de Decisão são modelos que realizam predições por meio de divisões recursivas dos dados, criando uma estrutura hierárquica baseada em regras. Esse tipo de modelo é capaz de capturar relações não lineares entre as variáveis e apresenta interpretabilidade fácil, embora possa apresentar tendência ao sobreajuste quando não adequadamente regularizado (BREIMAN et al., 1984).

- ***Random Forest*:** Método de *ensemble* baseado na combinação de múltiplas Árvores de Decisão treinadas a partir de subconjuntos aleatórios dos dados e das variáveis. A agregação de predições reduz a variância do modelo e melhora sua capacidade de generalização, sendo amplamente utilizada em problemas de regressão e classificação (BREIMAN, 2001).

- **KNN:** O algoritmo *K-Nearest* Neighbors é um método baseado em instâncias que realiza predições a partir da média dos valores associados aos k vizinhos mais próximos no espaço de atributos. Sua simplicidade e a ausência de fase explícita de treinamento tornam o KNN uma abordagem eficaz em determinados contextos, embora seu desempenho dependa fortemente da escolha de k e da escala dos dados (COVER; HART, 1967).

- **SVR:** A *Support Vector Regression* é uma adaptação do algoritmo de Support Vector Machines para problemas de regressão. O SVR busca encontrar uma função que minimize o erro dentro de uma margem definida, sendo capaz de modelar relações lineares e não lineares por meio do uso de funções kernel. Esse método é conhecido por

sua robustez em conjuntos de dados de média dimensão (SMOLA; SCHÖLKOPF, 2004).

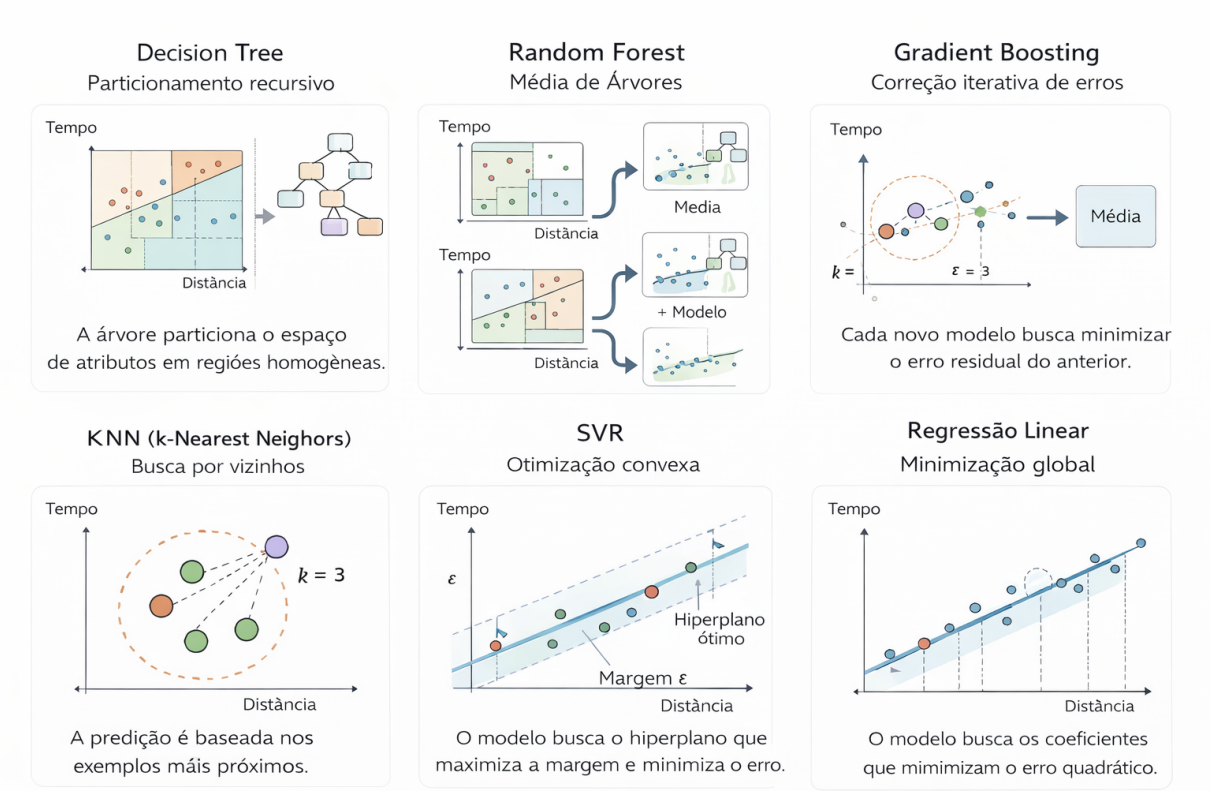

Figura 2.1: Ilustração conceitual dos mecanismos de busca dos algoritmos de aprendizado de máquina utilizados.

2.3.2 *Grid Search*

A seleção adequada de hiperparâmetros é um fator determinante para o desempenho de modelos de aprendizado de máquina, influenciando diretamente a capacidade de generalização. Nesse contexto, a técnica conhecida como *Grid Search* tem sido amplamente utilizada para otimizar esses valores de forma sistemática.

O *Grid Search* consiste em uma busca exaustiva por meio da avaliação de todas as combinações possíveis de um conjunto previamente definido de hiperparâmetros. Ao final do processo, a combinação com o melhor desempenho é selecionada como a configuração ideal (BERGSTRA; BENGIO, 2012).

Para garantir uma estimativa mais confiável do desempenho e reduzir o risco de *overfitting*[2], o processo de otimização pode ser conduzido em conjunto com a técnica

[2]Overfitting ocorre quando um modelo aprende padrões específicos do conjunto de treinamento, em

de validação cruzada (*cross-validation*), na qual o conjunto de dados é particionado em múltiplos subconjuntos (*folds*), permitindo que o modelo seja treinado e avaliado em diferentes divisões dos dados. Essa abordagem fornece uma avaliação mais robusta da capacidade de generalização do modelo (KOHAVI, 1995).

Essa técnica é amplamente empregada em problemas de regressão devido à sua facilidade de implementação e à garantia de uma avaliação completa do espaço de busca definido (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No presente trabalho, o *Grid Search* foi utilizado para otimizar os hiperparâmetros dos modelos de aprendizado de máquina empregados na predição do tempo de corrida, contribuindo para a obtenção de modelos mais precisos e robustos.

2.3.3 Avaliação e Métricas

Para avaliar o desempenho dos modelos de *Machine Learning*, foram utilizadas métricas amplamente empregadas em problemas de regressão. Essas métricas permitem quantificar o erro das predições e comparar o desempenho entre diferentes algoritmos.

A seguir, apresenta-se um resumo de cada métrica, acompanhado de sua respectiva equação. Onde y_i representa o valor observado, $\hat{y}_i$ o valor estimado pelo modelo, $\bar{y}$ a média dos valores observados e n o número total de observações.

- ***Mean Absolute Percentage Error* (MAPE):** Mede o erro percentual médio entre os valores reais e previstos. Essa métrica facilita a interpretação dos resultados em termos relativos, embora possa apresentar limitações quando os valores reais se aproximam de zero (HYNDMAN; KOEHLER, 2006).

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{2.1}$$

- ***Mean Absolute Error* (MAE):** Mede a média dos erros absolutos entre os valores reais e os valores previstos pelo modelo. Trata-se de uma métrica de fácil interpretação, pois expressa o erro médio na mesma unidade da variável-alvo (WILLMOTT; MATSUURA, 2005).

vez de generalizar para novos dados.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.2}$$

- ***Root Mean Squared Error* (RMSE):** O RMSE corresponde à raiz quadrada da média dos erros quadráticos. Essa métrica penaliza erros maiores de forma mais severa, sendo útil para identificar modelos que apresentam grandes desvios em suas predições (CHAI; DRAXLER, 2014).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.3}$$

- **Coeficiente de Determinação ($\mathbf{R^2}$):** O coeficiente de determinação R^2 indica a proporção da variância da variável dependente explicada pelo modelo. Valores mais próximos de 1 indicam melhor ajuste, enquanto valores próximos de 0 indicam baixo poder explicativo (NAGELKERKE, 1991).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.4}$$

3 Revisão Bibliográfica

Em Davidson et al. (2020), estudos que utilizam dados de *smartwatch* combinados com métodos de *machine learning* demonstraram que modelos como *Gradient Boosting, Regression Trees* podem estimar parâmetros de carga interna e esforço percebido (que são relacionados à performance de corrida) com MAPE baixo e correlação significativa com variáveis fisiológicas. Isto mostra a aplicabilidade de modelos de regressão e métricas de erro para avaliar padrões de esforço em atividades físicas monitoradas por *wearables*.

Em Nugroho (2025), pesquisas recentes em *big data* esportivo apresentam *pipelines* que coletam dados de *smartwatches* via API do Strava e aplicam modelos de *machine learning*, obtendo resultados com R^2 muito altos e MAPE baixo, demonstrando o potencial desses dados e algoritmos para predizer o desempenho de atletas.

Em Pirscoveanu e Oliveira (2023), foi demonstrado o potencial do uso de técnicas de aprendizado de máquina na análise de dados de corridas ao ar livre. Investigou-se a predição do esforço percebido instantâneo (*Rate of Perceived Exertion* – RPE) durante a corrida, utilizando variáveis biomecânicas e fisiológicas obtidas por acelerometria de *smartwatches*. Os autores aplicaram modelos de regressão, incluindo regressão linear, *Support Vector Regression* e regressão por processos Gaussianos, avaliando o desempenho dos modelos por meio Raiz do Erro Quadrático Médio (RMSE). Os resultados indicaram que modelos não lineares apresentaram melhor desempenho, especialmente quando informações específicas do indivíduo eram incorporadas ao treinamento.

O artigo Hong e Sun (2024) apresenta um estudo comparativo de modelos de aprendizado de máquina para estimar o consumo contínuo de oxigênio (VO) com base em dados de exercício físico obtidos por meio de teste de esteira (*treadmill*). Os autores treinaram e avaliaram modelos como Regressão Linear, *Gradient Boosting* e outras técnicas baseadas em árvores, utilizando métricas de desempenho amplamente reconhecidas, como R^2, MAE, RMSE e MAPE, para avaliar a precisão das previsões e a qualidade do ajuste dos modelos. A análise foi realizada com validação cruzada, garantindo que os resultados fossem robustos e generalizáveis a novos conjuntos de dados. Os resultados indicaram

que os modelos de *Gradient Boosting* apresentaram melhor desempenho, obtendo altos valores de R^2 e baixos erros absolutos, superando a regressão linear e outras abordagens tradicionais. O estudo destaca a eficácia de técnicas de *ensemble* em dados fisiológicos complexos, sugerindo que esses algoritmos podem ser aplicados ao monitoramento do desempenho físico e à otimização dos treinos de corrida.

Em relação a fatores que influenciam o desempenho em corridas, pode-se citar o estudo Beis et al. (2023), que analisou 668.509 corredores da Maratona de Berlim (1999–2019) para investigar como fatores climáticos influenciam o desempenho. Os principais fatores analisados foram temperatura, umidade, vento, precipitação, cobertura de nuvens e horas de sol. Os resultados mostraram que temperaturas mais altas e baixa umidade tendem a reduzir a velocidade média dos corredores, com efeito mais pronunciado nos homens. Outros fatores climáticos tiveram menor impacto. No geral, as condições climáticas explicaram cerca de 10% da variação no desempenho, indicando que outros fatores (preparo físico, estratégia, nutrição) são mais determinantes. O estudo também usou *machine learning* para identificar padrões complexos entre clima e desempenho. Os modelos confirmaram que temperatura e umidade são os fatores mais importantes, mas reforçaram que o clima, sozinho, não é suficiente para prever com precisão o tempo de corrida. Em resumo, o clima influencia a performance, mas precisa ser analisado junto com outros fatores, e técnicas de *machine learning* ajudam a compreender melhor essas relações complexas.

Em Weisz et al. (2024) investigou-se o uso de *Machine Learning* (ML) para prever a recuperação diária de atletas de *endurance*[3]. Foram coletados dados diários de 43 atletas ao longo de 12 semanas, incluindo o volume e a intensidade de treino, a duração e a qualidade do sono, a variabilidade da frequência cardíaca (HRV) e a dieta. Os resultados mostraram que os modelos de ML superaram as abordagens tradicionais, capturando relações não lineares entre as variáveis. Entre os preditores, as variáveis de sono se destacaram como fatores importantes, especialmente quando combinadas com outras métricas, indicando que a qualidade e a duração do sono influenciam significativamente a recu-

[3]Atletas de *endurance* são aqueles especializados em modalidades esportivas que exigem a manutenção de esforço físico prolongado, caracterizadas por elevada demanda de resistência cardiovascular, muscular e metabólica, como corrida de longa distância, ciclismo e triatlo.

peração. Além disso, o estudo ressaltou que a importância das *features* varia entre os atletas, o que evidencia que modelos personalizados são mais eficazes para entender como o sono e o treino afetam a recuperação. Embora o estudo não tenha medido diretamente o desempenho em corrida, ele mostra que uma boa recuperação, influenciada pelo sono e pelo treino, é fundamental para o desempenho em provas de *endurance* e que ML pode ajudar a quantificar e prever esses efeitos de forma individualizada.

Em Smith e Others (2025), 917 corredores, que participaram da Maratona de Boston em 2022, foram avaliados para compreender como o volume e a frequência de treino nos meses que antecedem a prova estão relacionados ao desempenho final (tempo de corrida) dos atletas. Os resultados mostraram que um maior volume semanal de corrida e mais sessões de treino de qualidade estavam associados a melhores tempos na maratona.

4 Materiais e Métodos

Neste capítulo, é apresentado, de forma detalhada, o processo metodológico adotado para o desenvolvimento e a implementação de uma ferramenta com suporte a modelos de *Machine Learning*, cujo objetivo é receber dados de corrida, treinar modelos preditivos e fornecer ao usuário uma estimativa do tempo necessário para percorrer uma determinada distância. O sistema foi concebido para auxiliar corredores amadores na análise de desempenho e no planejamento de seus treinos.

O desenvolvimento do trabalho foi dividido em quatro etapas, ilustradas na Figura 4.1. Na etapa de **Coleta de Dados**, as métricas de corrida foram obtidas a partir de duas fontes distintas. Os dados do primeiro voluntário foram coletados por meio de um *smartwatch*, utilizando o aplicativo Samsung Health, enquanto os do segundo voluntário foram obtidos por meio da exportação das atividades registradas na plataforma Strava. Essas fontes forneceram informações relevantes sobre o desempenho dos atletas, como distância percorrida, tempo de atividade, ritmo médio e outras métricas associadas à corrida.

Na etapa de **Pré-processamento dos Dados**, foram realizadas a limpeza, a organização e a normalização das informações coletadas, com o objetivo de garantir a consistência e a qualidade do conjunto de dados.

Em seguida, na etapa de **Treinamento dos modelos de *Machine Learning***, os dados processados foram utilizados para treinar os modelos de *Machine Learning*, possibilitando a identificação de padrões e a construção de modelos preditivos capazes de estimar o tempo necessário para a realização de uma corrida em uma distância definida pelo usuário.

Por fim, na etapa de **Predição de treinos**, a interface desenvolvida apresenta ao usuário as predições geradas pelo sistema, permitindo uma interação simples e intuitiva. Dessa forma, o sistema auxilia corredores amadores na compreensão de seu desempenho e no aprimoramento de seus treinos. Nas seções subsequentes, cada uma dessas etapas é descrita em detalhes, apresentando os procedimentos adotados e as decisões metodológicas

envolvidas em sua implementação.

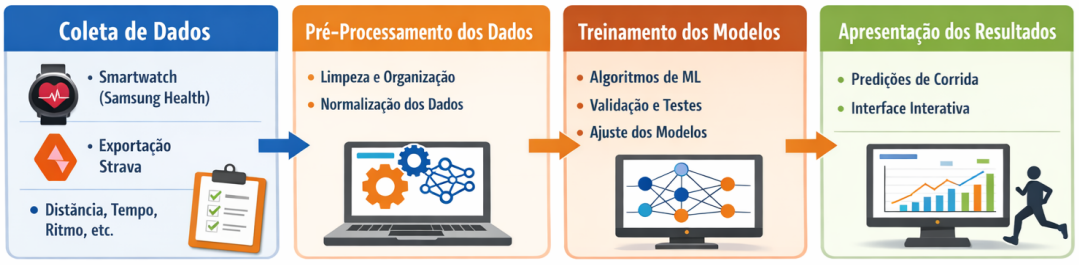

Figura 4.1: Fluxograma das etapas do desenvolvimento do sistema

4.1 Coleta de Dados

A etapa de Coleta de Dados consistiu no desenvolvimento de uma interface computacional simples e intuitiva, cujo objetivo é permitir a inserção dos dados de corrida utilizados pela ferramenta, como pode ser vista na Figura 4.2. Essa interface possibilita ao usuário realizar o envio de um arquivo compactado no formato .zip, contendo os registros de atividades físicas exportados a partir das plataformas Samsung Health e Strava. Tanto a exportação de dados do Samsung Health, realizada a partir de um *smartwatch* utilizado pelo voluntário 1, quanto a exportação de dados da plataforma Strava, referente ao voluntário 2, apresentam uma estrutura composta por múltiplos subarquivos no formato .csv. Esses subarquivos contêm informações relevantes no contexto da corrida, incluindo métricas como distância percorrida, tempo de atividade, ritmo médio, entre outros dados associados ao desempenho do atleta. Após o envio do arquivo .zip pela interface, o sistema realiza automaticamente a extração e a identificação dos subarquivos .csv correspondentes às atividades de corrida. Em seguida, os dados são lidos e organizados de acordo com sua origem, respeitando as particularidades estruturais de cada plataforma. Esse processo permite a unificação das informações provenientes de diferentes fontes, viabilizando sua posterior utilização nas etapas de processamento dos dados e treinamento dos modelos de *Machine Learning*. Dessa forma, a interface desenvolvida atua como um ponto central de entrada dos dados, garantindo flexibilidade quanto à origem das informações e facilitando a coleta de métricas essenciais para a construção do sistema de predição de desempenho em corrida.

Figura 4.2: Interface para envio dos dados exportados

4.2 Pré-processamento dos Dados

O pré-processamento dos dados teve como objetivo transformar os registros brutos provenientes das plataformas Samsung Health e Strava em um conjunto de dados consistente, padronizado e adequado para a etapa de modelagem preditiva. Embora os modelos tenham sido treinados separadamente para cada plataforma, adotou-se uma estratégia comum de padronização e limpeza dos dados, assegurando consistência metodológica e comparabilidade entre os conjuntos utilizados.

Inicialmente, os arquivos .csv exportados de ambas as plataformas foram carregados e submetidos a procedimentos de limpeza básica, incluindo a remoção de colunas redundantes, a eliminação de caracteres inválidos e a padronização dos nomes das variáveis. Em seguida, os registros foram filtrados para considerar exclusivamente as atividades de corrida, assegurando a coerência com o objetivo do estudo.

As informações temporais referentes ao início das atividades foram convertidas para o formato *datetime*, possibilitando a manipulação correta de datas e horários. A partir desses dados, foram extraídas, separadamente, a data e a hora de início de cada corrida, utilizadas tanto para a organização cronológica dos registros quanto para a integração com dados externos.

Para ambas as fontes, foram selecionadas métricas relevantes ao contexto da corrida, tais como distância percorrida, duração da atividade, velocidade máxima e tempo

em movimento, altitude mínima e máxima, inclinação mínima e máxima. A partir dessas variáveis, foram criadas novas métricas derivadas, incluindo o *pace* médio da corrida, calculado em minutos por quilômetro, e o volume acumulado de treino nos últimos sete dias, obtido por meio de uma janela temporal.

No que se refere às condições ambientais, os dados meteorológicos foram integrados de forma padronizada em ambas as plataformas. Sempre que as informações climáticas não estavam disponíveis ou apresentavam valores incompletos nos arquivos originais, foi realizada uma complementação automática por meio da consulta à API Open-Meteo (Open-Meteo, 2024), utilizando a data, o horário aproximado de início da atividade e as coordenadas geográficas associadas à corrida. Os dados foram obtidos em resolução horária, sendo selecionado o registro mais próximo ao horário de início da atividade para representar as condições ambientais enfrentadas pelo atleta.

No caso específico dos dados provenientes do Samsung Health, foi possível incorporar informações adicionais sobre o estado fisiológico do voluntário. Registros de sono foram processados e agregados por data, considerando métricas como a duração do sono, a pontuação geral, a recuperação física e a recuperação mental. Esses dados foram posteriormente integrados aos registros de corrida, permitindo a análise da influência do descanso no desempenho esportivo.

Após a integração das diferentes fontes de dados, foram realizados procedimentos adicionais de limpeza e transformação, incluindo a padronização dos formatos numéricos, a conversão de durações para segundos e a remoção de registros inconsistentes, como atividades com distância ou duração iguais ou inferiores a zero. Os valores ausentes foram tratados por meio da substituição pela média das variáveis numéricas correspondentes. A Tabela 4.1 apresenta as variáveis do Samsung Health que receberam imputação pela média e o número de registros afetados, enquanto a Tabela 4.2 apresenta a mesma análise para os dados do Strava.

Por fim, os conjuntos de dados foram ordenados cronologicamente e exportados em arquivos .csv, tornando-os aptos para utilização na etapa de treinamento e avaliação dos modelos de *Machine Learning* empregados no sistema proposto.

Tabela 4.1: Variáveis do Samsung Health que receberam imputação pela média

Variável	Quantidade de registros imputados
com.samsung.health.exercise.mean_speed	1
com.samsung.health.exercise.mean_heart_rate	25
sleep_duration	35
sleep_score	35
mental_recovery	35
com.samsung.health.exercise.mean_cadence	9
com.samsung.health.exercise.min_altitude	32
com.samsung.health.exercise.incline_distance	37
com.samsung.health.exercise.max_altitude	32
com.samsung.health.exercise.max_cadence	9
com.samsung.health.exercise.decline_distance	39
com.samsung.health.exercise.vo2_max	41
com.samsung.health.exercise.altitude_loss	54
com.samsung.health.exercise.altitude_gain	25

Tabela 4.2: Variáveis do Strava que receberam imputação pela média

Variável	Quantidade de registros imputados
Max Speed	10
Elevation Gain	10
Elevation Loss	29
Elevation High	23
Elevation Low	23
Max Grade	10

4.2.1 Descrição dos dados

Foram utilizados dois conjuntos de dados sobre atividades físicas. O conjunto do Samsung Health possui 171 registros, com maior detalhamento por atividade, incluindo métricas de desempenho físico, fisiológicas e de recuperação, como distância, velocidade média, duração, frequência cardíaca, sono, cadência, altitudes, ganho e perda de elevação, VO2 máximo, além de condições climáticas, volume semanal e *pace*.

Os dados analisados, gerados pelo Samsung Health, foram obtidos por meio da exportação de um conjunto de arquivos CSV e integrados com informações de clima, sono e exercícios. Esses arquivos fazem parte do arquivo ZIP exportado diretamente da ferramenta, contendo registros temporais de atividades físicas, métricas de sono e variáveis ambientais. A junção desses arquivos permitiu a construção de um único dataframe completo para análise, com informações sincronizadas entre os exercícios, os padrões de sono e as condições climáticas do dia da atividade.

Para cada variável, obtivemos as estatísticas descritivas, conforme apresentadas na Tabela 4.3.

Tabela 4.3: Estatísticas das variáveis - Samsung Health

Variável	Média	Desvio Padrão	Mínimo	25%	Mediana	75%	Máximo
com.samsung.health.exercise.distance(m)	6530.083	3416.395	30.546	5044.930	6243.085	8012.149	21272.14
com.samsung.health.exercise.mean_speed(m/s)	2.655226	0.403683	0.950165	2.508171	2.726691	2.864020	4.175581
com.samsung.health.exercise.duration(ms)	2425401	1149885	22799	1884795	2359326	2964090	7025916
com.samsung.health.exercise.mean_heart_rate(bpm)	147.9959	15.66953	74	147.9959	147.9959	153.25	198
sleep_duration(min)	447.2529	76.96145	83	447.2529	447.2529	471.75	599
temperature_x(°C)	19.78568	3.680149	11	17.2235	19.78568	21.89238	33
humidity_x(%)	84.50883	11.27012	43	79	84.50883	93	100
wind_speed_x(km/h)	7.557555	3.637482	0	5	7.557555	10	26
sleep_score	75.70115	8.164561	46	75.70115	75.70115	79	95
mental_recovery	78.10345	13.80143	33	78.10345	78.10345	87.25	98
volume_7d(m)	9299356	13272060	14874	379575	4463041	13836290	59650050
pace(min/km)	6.610230	1.693846	4.004726	5.809954	6.112957	6.648494	17.66081
com.samsung.health.exercise.mean_cadence(passos/min)	161.2215	18.67999	92.10859	160.0600	166.4930	172.4851	178.6522
com.samsung.health.exercise.min_altitude(m)	654.0145	232.0068	-51.67231	639.5683	724.4432	833.7413	882.657
com.samsung.health.exercise.incline_distance(m)	813.1951	481.6260	0	541.9790	796.5	939.8853	2750
com.samsung.health.exercise.max_altitude(m)	679.9748	235.5944	-42.59346	663.5278	761.0121	862.2840	909.081
com.samsung.health.exercise.max_cadence(passos/min)	179.7459	9.567645	107.119	176.6257	180	183.2234	223.475
com.samsung.health.exercise.decline_distance(m)	604.4001	327.6993	0	403.8015	604.4001	742.5	2222
com.samsung.health.exercise.vo2_max(ml/kg/min)	45.92159	1.557766	39.8	45.92159	45.92159	45.9525	50.02
com.samsung.health.exercise.altitude_loss(m)	90.54082	25.55804	17.038536	90.54082	90.54082	90.54082	280.4245
com.samsung.health.exercise.altitude_gain(m)	86.71312	71.29491	0	63.89252	86.71312	90.05163	860.779

Com base no cálculo das principais medidas estatísticas (média, desvio padrão, mínimo, percentis 25%, mediana, percentil 75% e máximo) para cada variável, observamos que:

- **Distâncias e duração** apresentam larga variabilidade, com média de aproximadamente 6,5 km percorridos por atividade e duração média próxima de 40 minutos (convertida em milissegundos).
- **Velocidade média** gira em torno de 2,6 m/s, compatível com ritmos de corrida leves a moderados.
- **Frequência cardíaca média** está próxima de 148 bpm, indicando intensidade moderada das atividades registradas.
- Variáveis climáticas indicam que as atividades ocorreram em faixas médias de **temperatura** de cerca de 20°C e de **umidade** em torno de 85%, com ventos leves a moderados.
- Indicadores de **sono** e **recuperação** sugerem que, em média, os participantes dormiram cerca de 447 minutos (7h27min), com ***scores* de sono** e de **recuperação mental** elevados, acima de 75.

- Medidas de **altitude**, **inclinação** e **cadência** indicam variações de terreno e de ritmo de movimento entre os registros.

- A métrica **volume_7d** apresenta grande amplitude, indicando variação no volume de treino semanal.

Essas estatísticas foram obtidas com base em dados limpos e pré-processados, garantindo que registros inválidos foram removidos e que valores ausentes foram devidamente imputados.

Já o conjunto do Strava apresenta 236 registros, com foco em métricas gerais de treino, como tempo de movimento, distância, velocidade máxima, ganho e perda de elevação, altitudes, inclinação, clima, volume semanal e *pace*. Dessa forma, os dados do Samsung Health fornecem informações mais detalhadas por registro, enquanto os do Strava apresentam maior número de registros, porém com menos variáveis por atividade.

Para cada variável, obtivemos a estatística descritiva, conforme mostrada na Tabela 4.4.

Tabela 4.4: Estatísticas das variáveis - Strava

Variável	**Média**	**Desvio Padrão**	**Mínimo**	**25%**	**Mediana**	**75%**	**Máximo**
Moving Time (s)	2832.81	2487.37	94.00	1510.00	2066.00	3036.00	17127.00
Distance.1 (m)	7388.41	6594.24	202.10	3931.00	5188.20	8000.00	42897.90
Max Speed (m/s)	3.6558	0.8061	1.511	3.147	3.629	3.995	8.490
Elevation Gain (m)	47.059	46.388	0.0	9.7	35.0	67.8	284.4
Elevation Loss (m)	223.531	221.680	0.0	112.0	204.0	268.0	2366.0
Max Grade (%)	7.155	8.054	0.0	3.5	5.8	7.155	49.1
temperature_x (°C)	21.474	3.908	10.236	19.007	21.474	23.807	32.207
humidity_x (%)	74.368	13.409	33.0	66.0	74.368	84.0	100.0
wind_speed_x (km/h)	6.455	3.741	0.509	3.827	5.805	7.594	24.575
volume_7d (m)	16888.93	14409.41	431.8	5363.4	11654.5	24707.2	69341.2
pace (mim/km)	6.613	1.128	3.016	5.813	6.358	7.300	11.051

Com base nos valores obtidos, observamos que:

- O **tempo em movimento (*Moving Time*)** varia bastante, desde atividades rápidas de poucos minutos (ex.: 94 s ou 1,5 min) até longos treinos que ultrapassam 17.000 s (4,7 h).

- A **distância percorrida (*Distance.1*)** também apresenta grande variabilidade, de menos de 1 km a mais de 40 km. Isso indica que o conjunto de dados abrange tanto treinos curtos quanto corridas de longa duração.

- A **velocidade máxima (*Max Speed*)** varia de aproximadamente 1,5 m/s a mais de 8 m/s, refletindo diferentes intensidades de esforço.

- O **pace médio** das atividades indica que grande parte das corridas apresenta ritmo entre 5 e 8 min/km, o que é compatível com corredores não profissionais.

- O **ganho e perda de elevação (*Elevation Gain / Loss*)** variam significativamente, de trechos planos (0 m) a percursos com mais de 2.000 m de variação altimétrica acumulada.

- O **gradiente máximo (*Max Grade*)** das atividades varia de 0% (plano) a quase 50%, indicando que algumas corridas ocorreram em subidas mais íngremes.

- A **temperatura média** durante as atividades está entre 10–28 °C.

- A **umidade relativa** varia de 33% a 100% e a **velocidade do vento** de 0 a 25 m/s, indicando uma diversidade de condições meteorológicas.

- A coluna **volume_7d** registra o total de metros percorridos na semana, com valores que variam de menos de 1 km a mais de 55 km, evidenciando diferentes cargas de treino.

Há atividades curtas e longas, planas e com elevação acentuada, o que permite análises de desempenho em diferentes contextos. A variação de todas as métricas sugere que o conjunto de dados é adequado para estudos de desempenho, modelagem preditiva de tempo/distância e análise de esforço físico.

As Tabelas 4.5 e 4.6 apresentam exemplos das primeiras linhas do dataframe final obtido a partir dos dados coletados pelo Samsung Health, após o processo de pré-processamento, e utilizado no treinamento dos modelos de Machine Learning. De forma complementar, as Tabelas 4.7 e 4.8 apresentam as primeiras linhas do dataframe final, construído com dados provenientes do Strava, após as etapas de tratamento e preparação dos dados para o treinamento dos modelos. Essas tabelas permitem visualizar a estrutura dos dados processados e evidenciam a diversidade de métricas disponíveis em cada plataforma.

Tabela 4.5: Exemplo das primeiras linhas do dataframe final do Samsung Health — variáveis de corrida, fisiologia e percurso

Distance (m)	Mean Speed (m/s)	Duration (ms)	Mean HR (bpm)	Pace (min/km)	Mean Cad. (spm)	Min Alt (m)	Max Alt (m)	Alt Gain (m)	Alt Loss (m)	Incline Dist (m)	Decline Dist (m)	VO2 Max (ml/kg/min)
5000.0	2.083	2400000	147.72	8.00	160.29	654.89	680.67	86.04	89.19	807.13	599.82	45.92
5444.71	2.539	2144710	76.00	6.57	158.55	845.69	872.81	86.04	58.63	695.71	354.35	45.92
5071.89	3.244	1563261	175.00	5.14	170.91	-18.90	-14.07	0.00	89.19	6.74	599.82	43.84
6544.71	2.605	2511967	158.00	6.40	163.59	836.16	862.26	66.00	115.00	1128.76	542.53	44.35
3013.41	1.806	1668829	119.00	9.23	131.37	797.89	822.28	25.26	89.19	359.47	225.63	39.80
5366.08	2.743	1956220	153.00	6.08	166.21	19.94	24.17	0.00	89.19	6.92	599.82	40.31
5389.26	2.636	2044755	152.00	6.32	161.89	-8.45	-3.88	23.91	89.19	807.13	1.76	40.39
7106.96	2.827	2514204	162.00	5.90	164.86	837.12	862.36	860.78	89.19	1086.74	421.01	40.90
3970.26	3.151	1260000	147.72	5.29	160.29	654.89	680.67	86.04	72.15	807.13	599.82	45.92

Tabela 4.6: Exemplo das primeiras linhas do dataframe final do Samsung Health — variáveis de sono, clima e carga

Sleep Dur (min)	Sleep Score (score)	Mental Recov (score)	Temp (°C)	Humidity (%)	Wind (km/h)	Volume 7d (m)	Max Cadence (spm)
552.0	62.0	43.0	22.56	75.0	15.14	7231478	179.71
506.0	82.0	75.0	23.00	66.0	7.90	5444707	172.71
506.0	84.0	72.0	20.71	56.0	13.01	5071894	180.91
567.0	84.0	87.0	19.00	81.0	8.00	6544712	174.57
492.0	80.0	92.0	22.00	80.0	3.00	3013412	170.12
380.0	71.0	71.0	21.86	89.0	6.92	8379495	172.68
412.0	84.0	86.0	29.00	60.0	18.00	13768758	170.64
417.0	65.0	81.0	26.00	66.0	8.00	7106958	173.32
483.0	82.0	89.0	23.00	72.0	11.00	397026	179.71

Tabela 4.7: Exemplo das primeiras linhas do dataframe final do Strava — variáveis de corrida e percurso

Moving Time (s)	Distance (m)	Max Speed (m/s)	Elev. Gain (m)	Elev. Loss (m)	Elev. High (m)	Elev. Low (m)	Max Grade (%)
2514.0	6627.9	3.37	39.3	236.0	868.0	860.5	8.1
17127.0	42897.9	7.091	231.6	2366.0	44.2	-3.5	49.1
938.0	2042.8	2.695	9.7	34.0	862.9	853.2	3.5
1953.0	4044.6	2.98	20.1	197.0	867.9	860.5	4.3
1943.0	3931.0	3.286	41.1	229.0	870.5	844.1	12.4
1986.0	4059.6	2.987	35.0	146.0	865.9	844.1	6.5
1596.0	3544.6	4.394	21.1	89.0	751.0	737.0	8.3
2269.0	4466.2	2.747	12.4	114.0	860.8	848.2	3.4
2364.0	4966.8	2.852	30.9	170.0	866.2	853.8	7.5

Tabela 4.8: Exemplo das primeiras linhas do dataframe final do Strava — variáveis ambientais e derivadas

Temperature (°C)	Humidity (%)	Wind Speed (km/h)	Volume_7d (m)	Pace (min/km)
18.26	57.0	11.79	15124.3	6.32
14.46	95.0	2.74	55013.6	6.65
21.96	63.0	13.91	2042.8	7.65
23.66	62.0	10.44	6087.4	8.05
23.21	61.0	14.59	7975.6	8.24
19.56	62.0	2.52	4059.6	8.15
22.76	66.0	13.90	7604.2	7.50
18.26	88.0	3.22	8010.8	8.47
21.56	68.0	10.31	9433.0	7.93

4.2.2 Matriz de Correlação

A matriz de correlação é uma ferramenta estatística que permite avaliar a relação linear entre pares de variáveis em um conjunto de dados. Cada elemento da matriz representa o coeficiente de correlação, que varia entre -1 e 1, indicando a força e a direção da relação. Valores próximos de 1 ou -1 indicam correlações fortes, positivas ou negativas, respectivamente, enquanto valores próximos de zero indicam pouca ou nenhuma associação linear entre as variáveis analisadas. Essa matriz é amplamente utilizada para identificar padrões, relações e possíveis redundâncias entre características, facilitando a interpretação e seleção de variáveis em estudos científicos.

A Figura 4.3 apresenta a matriz de correlação contendo as dez variáveis com maior relevância dentre aquelas coletadas pelo Samsung Health. Ressalta-se que os nomes das variáveis foram ajustados para melhorar a legibilidade da visualização, sem comprometer o entendimento ou o significado original das informações apresentadas.

Primeiramente, destaca-se a forte correlação positiva (r = 0,98) entre a variável distância do exercício e a duração do exercício. Esse resultado é coerente, pois atividades físicas de maior duração tendem a cobrir maiores distâncias, especialmente em modalidades contínuas, como a corrida. Da mesma forma, a velocidade média apresenta uma correlação positiva moderada (r = 0,58) com a cadência média, indicando que um aumento no número de passos ou ciclos por minuto está associado a um aumento na velocidade média da atividade.

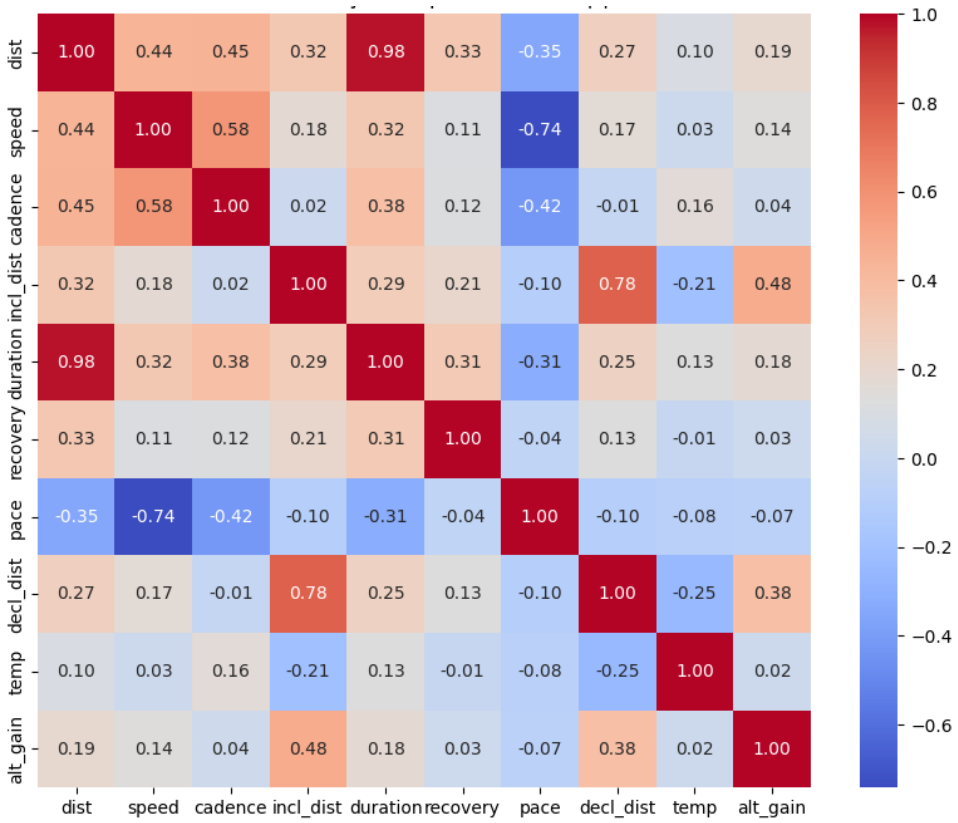

Figura 4.3: Matriz de correlação - Samsung

Outra relação importante está entre as variáveis de distância inclinada e distância declinada (r = 0,78), o que sugere que os percursos analisados apresentam variações de terreno que incluem tanto subidas quanto descidas, comportamento comum em rotas reais de treino e competição.

Em relação ao ritmo da atividade (*pace*), verifica-se correlação negativa significativa com a velocidade média (r = -0,74) e com a cadência média (r = -0,42). Tal relação é esperada, pois o pace representa o tempo necessário para percorrer uma distância fixa, sendo inversamente proporcional à velocidade. Assim, uma velocidade maior implica um ritmo menor (tempo menor para a mesma distância), o que explica a correlação negativa observada.

A variável ganho de altitude está positivamente correlacionada com a distância inclinada (r = 0,48), o que confirma a lógica de que as subidas contribuem diretamente para o aumento do ganho de altitude durante o exercício.

Por outro lado, variáveis como recuperação mental e temperatura apresentaram baixas correlações com as demais características, indicando que seus efeitos diretos sobre

as métricas de exercício avaliadas são pouco expressivos no conjunto de dados analisado. Na literatura, a recuperação mental é um aspecto relacionado ao bem-estar geral e pode influenciar indiretamente o desempenho, mas essa relação não se mostrou forte nos dados aqui apresentados. A influência da temperatura, por sua vez, pode variar conforme a faixa térmica e o nível de adaptação dos praticantes, o que justifica a baixa correlação observada.

Em resumo, as correlações identificadas na matriz refletem de forma consistente os aspectos fisiológicos e mecânicos conhecidos no exercício físico, evidenciando as relações esperadas entre distância, duração, velocidade, ritmo e variações de terreno. Aspectos psicofisiológicos e ambientais, como a recuperação mental e a temperatura, apresentaram menor associação direta, indicando a necessidade de análises complementares para uma melhor compreensão de seus impactos.

De forma análoga, a Figura 4.4 apresenta a matriz de correlação contendo as dez variáveis com maior relevância provenientes do Strava, na qual se observa, inicialmente, uma correlação extremamente forte entre *Moving Time* e *Distance* (r = 0,99), o que é consistente com a literatura sobre esportes de *endurance*, uma vez que a distância percorrida é diretamente determinada pelo tempo em movimento, especialmente quando a variabilidade de velocidade média entre sessões é limitada.

As variáveis relacionadas à altimetria (*Elevation Loss* e *Elevation Gain*) apresentaram correlações fortes a moderadas com *Moving Time* (r = 0,85) e *Distance* (r = 0,83 e r = 0,41, respectivamente). Esses resultados são coerentes com estudos que demonstram que percursos mais longos tendem a acumular maior ganho e perda de elevação, particularmente em ambientes não urbanos ou com topografia variável. A correlação moderada entre *Elevation Gain* e *Elevation Loss* (r = 0,53) sugere que, embora relacionadas, essas variáveis não são perfeitamente simétricas, o que pode refletir diferenças no ponto inicial e final do percurso, bem como limitações inerentes à precisão da altimetria baseada em GPS.

O volume acumulado de treinamento em sete dias (volume_7d) apresentou correlação moderada a forte com *Distance* (r = 0,63) e *Moving Time* (r = 0,61), indicando que o aumento do volume está associado a sessões mais longas e maior tempo total de

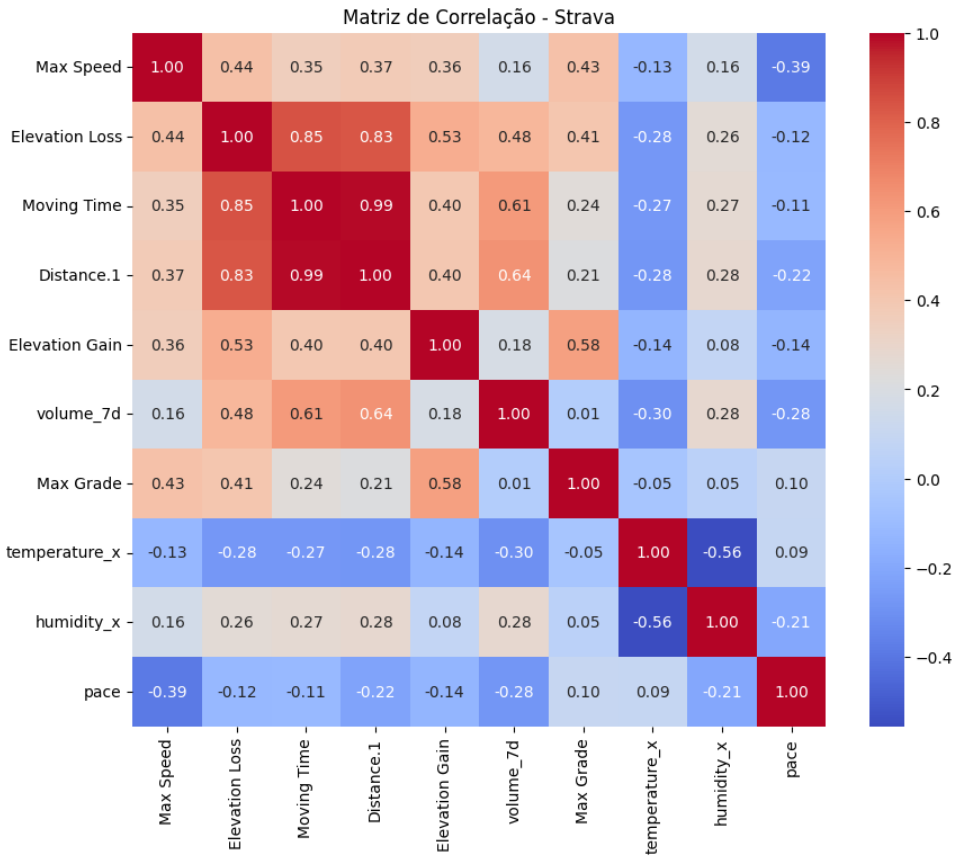

Figura 4.4: Matriz de correlação - Strava

exposição ao exercício. Esse achado está alinhado com conceitos discutidos na literatura sobre carga e sua relação com desempenho.

A variável *Max Speed* demonstrou correlações fracas a moderadas com *Distance* (r = 0,39), *Elevation Loss* (r = 0,44) e *Elevation Gain* (r = 0,37). Esses valores sugerem que a velocidade máxima atingida durante uma sessão é parcialmente influenciada pelo comprimento e pela topografia do percurso, especialmente pela presença de trechos descendentes, porém permanece amplamente dependente de fatores contextuais, como características técnicas do terreno e intenção do atleta. A literatura indica que a velocidade máxima não constitui um indicador robusto da intensidade global em esportes de endurance, reforçando a interpretação desses resultados.

A inclinação máxima (*Max Grade*) apresentou correlação praticamente nula com o volume semanal (volume_7d), evidenciando que essa variável representa uma característica pontual do percurso, independente da carga acumulada de treinamento. Tal comportamento é esperado, uma vez que a inclinação máxima reflete um evento específico e não a exposição total ao esforço.

Em relação às condições ambientais, a temperatura (temperature_x) apresentou correlações negativas fracas a moderadas com Distance, Moving Time e volume_7d (r entre -0,28 e -0,31). Esses resultados estão em consonância com evidências de que temperaturas mais elevadas estão associadas à redução do volume e da duração do exercício, em função do aumento do estresse térmico e cardiovascular.

Em síntese, a matriz de correlação revela um conjunto de relações amplamente coerentes com a literatura sobre carga externa, fisiologia do exercício e análise de treinamento, reforçando a validade das variáveis analisadas.

Mesmo após a análise das correlações entre as *features*, optou-se por manter todas as variáveis detalhadas no treinamento do modelo, considerando que cada uma pode contribuir de forma complementar para a predição do desempenho, além de evitar a perda de informações.

4.3 Treinamento dos modelos de *Machine Learning*

A etapa de treinamento dos modelos de *Machine Learning* teve como objetivo construir e avaliar modelos capazes de prever o tempo de duração de uma corrida a partir de métricas de desempenho e de contexto previamente pré-processadas. Para isso, foi definido como variável-alvo (*target*) o tempo total de duração da atividade, enquanto as demais variáveis compuseram o conjunto de atributos de entrada do modelo.

Como os dados foram coletados a partir de duas fontes distintas, Samsung Health e Strava, foram realizados dois treinamentos separados, utilizando dataframes específicos para cada plataforma. Dessa forma, cada conjunto de modelos foi ajustado e avaliado individualmente, considerando as características e métricas disponíveis em cada fonte de dados.

Inicialmente, o conjunto de dados foi dividido em variáveis independentes e variável dependente. Em seguida, os dados foram particionados em conjuntos de treinamento e teste, utilizando a proporção de 80% para treinamento e 20% para teste, com o uso de uma semente aleatória fixa, garantindo a reprodutibilidade dos experimentos. Os conjuntos resultantes foram armazenados em arquivos separados, possibilitando análises posteriores e validações adicionais.

Com o objetivo de avaliar diferentes abordagens de modelagem, foram selecionados seis algoritmos de regressão amplamente utilizados em problemas de predição contínua: Regressão Linear, *Support Vector Regression* (SVR), *K-Nearest Neighbors* (KNN), Árvore de Decisão, *Random Forest* e *Gradient Boosting*. Para cada modelo, foi construída uma *pipeline* contendo uma etapa de padronização dos dados por meio do *StandardScaler*, seguida pelo algoritmo de regressão propriamente dito. Essa abordagem garante que os atributos de entrada estejam na mesma escala, o que é especialmente relevante para modelos sensíveis à magnitude dos dados.

O processo de treinamento foi realizado com o auxílio da técnica de *Grid Search* associada à validação cruzada (*cross-validation*) com cinco *folds* (k = 5). Para cada algoritmo, foram definidos conjuntos de hiperparâmetros, a serem testados, permitindo a identificação da configuração que minimiza o erro de previsão.

A Tabela 4.9 apresenta os hiperparâmetros iniciais considerados pelo *Grid Search* para cada modelo avaliado, sendo utilizado o mesmo espaço de busca tanto para os dados provenientes do Samsung Health quanto para a plataforma Strava.

Tabela 4.9: Espaço de busca dos hiperparâmetros definido para o Grid Search

Modelo	**Parâmetros iniciais avaliados**
SVR	$C \in \{0.1, 1, 10\}$; $\epsilon \in \{0.01, 0.1\}$; kernel $\in \{\text{rbf}, \text{linear}\}$
Random Forest	n_estimators $\in \{100, 200\}$; max_depth $\in \{\text{None}, 10, 20\}$; min_samples_split $\in \{2, 5\}$; min_samples_leaf $\in \{1, 2\}$
KNN	n_neighbors $\in \{3, 5, 7\}$; weights $\in \{\text{uniform}, \text{distance}\}$
Linear Regression	Não aplicável (modelo sem hiperparâmetros ajustáveis)
Gradient Boosting	n_estimators $\in \{100, 200\}$; learning_rate $\in \{0.05, 0.1\}$; max_depth $\in \{3, 5\}$
Decision Tree	max_depth $\in \{\text{None}, 5, 10\}$; min_samples_split $\in \{2, 5\}$

As Tabelas 4.10 e 4.11 apresentam os melhores hiperparâmetros selecionados pela técnica de *Grid Search* para os modelos treinados com dados provenientes do Samsung Health e do Strava, respectivamente. Observa-se que, embora alguns modelos tenham apresentado configurações semelhantes, como o SVR e o *Gradient Boosting*, outros algoritmos, como *Random Forest* e *Decision Tree*, apresentaram diferenças nos hiperparâmetros ótimos em função da fonte de dados utilizada (Tabelas 4.10 e 4.11).

Tabela 4.10: Melhores hiperparâmetros obtidos pelo Grid Search para os dados do Samsung Health

Modelo	**Hiperparâmetros selecionados**
SVR	$C = 10$, $\epsilon = 0.01$, kernel = linear
Random Forest	n_estimators = 200; max_depth = None; min_samples_split = 2; min_samples_leaf = 1
KNN	n_neighbors = 3; weights = distance
Linear Regression	Não possui hiperparâmetros ajustáveis
Gradient Boosting	n_estimators = 200; learning_rate = 0.05; max_depth = 3
Decision Tree	max_depth = None; min_samples_split = 2

Tabela 4.11: Melhores hiperparâmetros obtidos pelo Grid Search para os dados do Strava

Modelo	**Hiperparâmetros selecionados**
SVR	$C = 10$, $\epsilon = 0.01$, kernel = linear
Random Forest	n_estimators = 200; max_depth = 10; min_samples_split = 2; min_samples_leaf = 1
KNN	n_neighbors = 3; weights = distance
Linear Regression	Não possui hiperparâmetros ajustáveis
Gradient Boosting	n_estimators = 200; learning_rate = 0.05; max_depth = 3
Decision Tree	max_depth = 10; min_samples_split = 2

Após a identificação do melhor conjunto de hiperparâmetros para cada modelo, o desempenho foi avaliado no conjunto de teste previamente separado. As predições geradas foram comparadas aos valores reais de duração das corridas, sendo calculadas métricas quantitativas de desempenho, incluindo o Erro Médio Absoluto (MAE), a Raiz do Erro Quadrático Médio (RMSE), o coeficiente de determinação (R^2) e o Erro Percentual Médio Absoluto (MAPE). Essas métricas permitiram uma análise abrangente da precisão e da capacidade de generalização dos modelos treinados.

Os modelos treinados, juntamente com seus respectivos hiperparâmetros otimizados, foram então serializados e armazenados em arquivos, possibilitando sua posterior reutilização na interface desenvolvida para realização de predições em tempo de execução. Adicionalmente, foi realizada uma análise dos erros absolutos das predições no conjunto de teste, identificando os casos com maior discrepância entre os valores reais e previstos, contribuindo para uma avaliação qualitativa do comportamento dos modelos.

Ao final desse processo, os resultados obtidos por cada algoritmo foram organiza-

dos em formato tabular, permitindo a comparação direta entre os modelos e subsidiando a escolha daquele mais adequado para integração ao sistema proposto.

4.3.1 Avaliação dos Modelos

A Tabela 4.12 apresenta o desempenho dos modelos de *Machine Learning* empregados na previsão do tempo de corrida a partir de dados do Samsung Health. De modo geral, observa-se que os modelos baseados em árvores e métodos de ensemble apresentaram desempenho superior em relação aos modelos lineares e baseados em distância, indicando a presença de relações não lineares e interações complexas entre as variáveis explicativas e o tempo de corrida.

Tabela 4.12: Desempenho dos modelos de Machine Learning na previsão de tempo de corrida - Samsung Health

Modelo	**MAE**	**RMSE**	**R^2**	**MAPE (%)**
Gradient Boosting	**2.62**	**5.86**	**0.93**	**6.37**
Random Forest	3.34	6.61	0.92	8.50
Decision Tree	4.03	7.34	0.90	10.57
Linear Regression	3.05	7.87	0.88	32.40
KNN	9.16	15.86	0.51	21.59
SVR	15.78	22.94	-0.02	69.31

O modelo *Gradient Boosting* obteve o melhor desempenho global, apresentando o menor MAE (2,62), o menor RMSE (5,86), elevado coeficiente de determinação (R^2 = 0,93) e baixo MAPE (6,37%). Esses resultados indicam alta precisão preditiva e excelente capacidade de explicação da variabilidade do tempo de corrida. A superioridade desse modelo é consistente com a literatura, que aponta o *Gradient Boosting* como particularmente eficaz em cenários com dados tabulares heterogêneos e relações não lineares, comuns em dados fisiológicos e de treinamento esportivo.

O *Random Forest* apresentou desempenho semelhante, embora ligeiramente inferior ao *Gradient Boosting*, com MAE de 3,34, RMSE de 6,61 e R^2 de 0,92. O bom desempenho desse modelo pode ser atribuído à sua capacidade de reduzir variância por meio da agregação de múltiplas árvores de decisão, tornando-o robusto a ruídos e outliers. Ainda assim, o *Random Forest* tende a apresentar menor capacidade de ajuste fino em comparação ao *Gradient Boosting*, o que pode explicar a diferença observada nas métricas

de erro.

O modelo de *Decision Tree* isolado apresentou desempenho inferior aos métodos de ensemble, com MAE de 4,03 e R^2 de 0,90. Embora ainda apresente resultados razoáveis, esse comportamento é esperado, uma vez que árvores individuais são mais suscetíveis a *overfitting* e possuem menor capacidade de generalização, conforme amplamente discutido na literatura de aprendizado de máquina.

A Regressão Linear apresentou desempenho intermediário, com MAE de 3,05 e R^2 de 0,88. Apesar de um erro absoluto relativamente baixo, o valor elevado de MAPE (32,40%) sugere limitações na modelagem de variações proporcionais no tempo de corrida, especialmente em valores mais baixos. Esse resultado indica que a relação entre as variáveis independentes e o tempo de corrida não é estritamente linear, o que reduz a adequação desse modelo para o problema em questão.

Os modelos KNN e SVR apresentaram desempenho substancialmente inferior. O KNN obteve MAE de 9,16 e R^2 de 0,51, indicando baixa capacidade explicativa e elevada sensibilidade à escolha de métricas de distância e à escala das variáveis. Já o SVR apresentou o pior desempenho entre os modelos avaliados, com MAE de 15,78, RMSE de 22,94 e R^2 negativo (-0,02), sugerindo que o modelo foi incapaz de capturar padrões relevantes nos dados, performando pior do que uma predição baseada na média. Esses resultados podem estar associados à necessidade de maior ajuste de hiperparâmetros e à sensibilidade do SVR à dimensionalidade e à distribuição das variáveis.

Tabela 4.13: Desempenho dos modelos de *Machine Learning* na previsão de tempo de corrida - Strava

Modelo	**MAE**	**RMSE**	**R^2**	**MAPE (%)**
Gradient Boosting	**1.65**	**2.49**	**0.99**	**6.14**
Linear Regression	3.61	5.19	0.98	10.52
Random Forest	3.04	6.37	0.98	7.50
Decision Tree	3.73	5.26	0.98	11.10
SVR	8.94	16.23	0.85	24.84
KNN	10.26	14.80	0.87	33.73

A Tabela 4.13 apresenta o desempenho dos modelos de *Machine Learning* na previsão do tempo de corrida a partir dos dados do Strava. De forma geral, observa-se desempenho superior em comparação aos modelos treinados com dados do Samsung Health, sugerindo maior consistência e qualidade das variáveis disponíveis nessa plataforma. O

modelo *Gradient Boosting* apresentou desempenho substancialmente superior aos demais, com MAE de 1,65, RMSE de 2,49, R^2 igual a 0.99 e MAPE de 6,14%. Esses resultados indicam capacidade de explicação da variabilidade do tempo de corrida e elevada precisão preditiva. Tal desempenho reforça a adequação de métodos de *boosting* para dados esportivos tabulares, nos quais interações complexas entre variáveis como distância, altimetria, tempo em movimento e condições ambientais desempenham papel determinante no desempenho do atleta.

A Regressão Linear apresentou desempenho consistente, com R^2 de 0,98 e MAE de 3,61, indicando que grande parte da variabilidade do tempo de corrida pode ser explicada por relações aproximadamente lineares nos dados do Strava. Ainda assim, o erro percentual (MAPE de 10,52%) foi superior ao observado no *Gradient Boosting*, sugerindo limitações na captura de efeitos não lineares e interações entre variáveis, conforme amplamente discutido na literatura de modelagem preditiva.

O *Random Forest* e o *Decision Tree* apresentaram desempenhos semelhantes, ambos com R^2 de 0,98. O *Random Forest* obteve MAE inferior (3,04) e MAPE mais baixo (7,50%) em comparação à árvore de decisão isolada, confirmando a vantagem dos métodos de *ensemble* na redução da variância e melhoria da generalização. A *Decision Tree*, por sua vez, apresentou desempenho razoável, porém com maior erro médio, refletindo sua maior suscetibilidade a ajustes excessivos aos dados de treinamento.

Os modelos SVR e KNN apresentaram desempenho inferior em relação aos demais, com MAE elevados (8,94 e 10,26, respectivamente) e valores de R^2 entre 0,85 e 0,87. Esses resultados indicam menor capacidade de generalização e sensibilidade à escolha de hiperparâmetros, bem como à escala e distribuição das variáveis. Apesar de apresentarem valores de R^2 relativamente altos, os erros absolutos e percentuais elevados sugerem limitações práticas desses modelos para aplicações preditivas precisas no contexto analisado.

De maneira geral, os resultados obtidos com os dados do Strava evidenciam que modelos baseados em árvores e métodos de *ensemble*, especialmente o *Gradient Boosting*, são os mais adequados para a previsão do tempo de corrida. Além disso, o desempenho significativamente superior em relação aos dados do Samsung Health sugere que o Strava

fornece variáveis mais diretamente relacionadas ao desempenho, como métricas detalhadas de altimetria e tempo em movimento, o que potencializa a capacidade preditiva dos modelos.

As Tabelas 4.14 e 4.15 mostram os tempos estimados pelos diferentes modelos de *Machine Learning* para distâncias entre 5 km e 30 km, permitindo avaliar não apenas o desempenho médio dos modelos, mas também sua coerência fisiológica, consistência ao longo das distâncias e capacidade de extrapolação. De modo geral, observa-se que os modelos treinados com dados do Strava apresentam estimativas mais plausíveis e progressivas com o aumento da distância, especialmente nos modelos baseados em árvores e métodos de *ensemble*. Em contraste, os modelos ajustados com dados do Samsung Health exibem maior variabilidade e, em alguns casos, estimativas fisiologicamente implausíveis, sobretudo em distâncias mais longas.

Tabela 4.14: Tempo estimado pelos modelos para 5 km, 8 km e 10 km

Modelo	**5 km**		**8 km**		**10 km**	
	Strava	**Sams.**	**Strava**	**Sams.**	**Strava**	**Sams.**
Gradient Boosting	26m17s	33m09s	42m19s	50m03s	54m47s	58m25s
Random Forest	28m45s	32m27s	44m33s	49m55s	59m43s	55m55s
Linear Regression	43m45s	53m39s	1h02m	71m23s	1h14m10s	1h23m12s
Decision Tree	28m07s	30m25s	46m35s	49m54s	1h02m17s	50m17s
KNN	33m13s	19m30s	33m14s	19m31s	33m15s	19m31s
SVR	25m27s	38m53s	32m19s	38m54s	36m53s	38m54s

Tabela 4.15: Tempo estimado pelos modelos para 15 km, 20 km e 30 km

Modelo	**15 km**		**20 km**		**30 km**	
	Strava	**Sams.**	**Strava**	**Sams.**	**Strava**	**Sams.**
Gradient Boosting	1h20m54s	1h24m19s	1h48m25s	1h27m21s	3h28m56s	1h27m21s
Random Forest	1h28m45s	1h22m21s	2h06m18s	1h24m02s	3h05m03s	1h24m02s
Linear Regression	1h44m35s	1h52m46s	2h15m00s	2h22m19s	3h15m50s	3h21m26s
Decision Tree	1h25m39s	1h20m24s	1h42m06s	1h20m24s	3h12m19s	1h20m24s
KNN	33m16s	19m31s	33m18s	19m31s	33m21s	19m32s
SVR	48m18s	38m56s	59m44s	38m57s	1h22m34s	39m00s

Como pode ser observado nas Tabelas 4.14 e 4.15, o *Gradient Boosting* demonstrou comportamento consistente para as distâncias de 5 km a 20 km em ambas as plataformas, com aumento progressivo do tempo estimado conforme a distância cresce, respeitando padrões esperados de desempenho em corrida. Para os dados do Strava, os

tempos estimados mantêm uma relação aproximadamente linear com a distância, refletindo boa capacidade de generalização. No entanto, para 30 km, observa-se uma superestimação relevante no Strava e uma subestimação acentuada no Samsung Health, sugerindo limitações do modelo na extrapolação para distâncias menos representadas no conjunto de treinamento.

Os modelos *Random Forest* e *Decision Tree* apresentaram estimativas razoáveis para distâncias curtas e intermediárias (5–15 km), especialmente com dados do Strava. Entretanto, para distâncias mais longas, observa-se aumento da variabilidade e, em alguns casos, subestimação do tempo de corrida nos dados do Samsung Health, indicando possível overfitting ou sensibilidade excessiva a padrões locais presentes nos dados de treino.

A Regressão Linear apresentou comportamento monotônico e previsível em todas as distâncias, com aumento contínuo do tempo estimado à medida que a distância cresce. Embora esse padrão seja desejável do ponto de vista fisiológico, os tempos estimados tendem a ser sistematicamente mais elevados do que os observados nos modelos não lineares, sugerindo que a suposição de linearidade não captura adequadamente variações de ritmo, fadiga e influência do terreno, especialmente em distâncias menores.

Os modelos KNN e SVR apresentaram estimativas inconsistentes ao longo de todas as distâncias analisadas. Em particular, o KNN manteve praticamente o mesmo tempo estimado independentemente da distância, tanto para o Strava quanto para o Samsung Health, o que é fisiologicamente inviável e indica falha clara na capacidade de generalização. Esse comportamento sugere que o modelo está excessivamente dependente de vizinhos próximos no espaço de características, sem capturar a relação estrutural entre distância e tempo.

O SVR também apresentou subestimações severas, sobretudo para distâncias longas (20 km e 30 km), com tempos incompatíveis com o desempenho humano esperado. Esses resultados reforçam as limitações desse modelo no contexto analisado, possivelmente associadas à sensibilidade à escolha de hiperparâmetros, além da dificuldade em extrapolar para regiões fora da distribuição original dos dados.

Um aspecto fundamental para a interpretação dos resultados dos modelos refere-se ao intervalo de distâncias presente nos conjuntos de dados. Os dados do Samsung

Health contemplam atividades com distâncias máximas de até 21.272,14 m, enquanto os dados do Strava incluem atividades significativamente mais longas, alcançando 42.897,9 m. Essa diferença estrutural tem implicações diretas na capacidade de aprendizado e extrapolação dos modelos.

Nos modelos treinados com dados do Samsung Health, as estimativas para distâncias superiores ao limite observado no conjunto de treinamento (por exemplo, 30 km) configuram extrapolações fora do domínio dos dados, o que explica as previsões inconsistentes e fisiologicamente implausíveis observadas em vários algoritmos, especialmente nos modelos baseados em vizinhança (KNN) e margens (SVR). Esses modelos dependem fortemente da distribuição original dos dados e tendem a falhar quando aplicados a valores não previamente observados.

Por outro lado, os modelos treinados com dados do Strava apresentaram maior estabilidade nas estimativas para distâncias longas, uma vez que o conjunto de treinamento já inclui corridas de longa duração, permitindo que os algoritmos aprendam padrões associados à fadiga acumulada e à redução progressiva do ritmo. Isso explica a maior coerência e precisão observadas nas previsões para 20 km e 30 km nessa base de dados.

4.4 Predição de treinos

A etapa de predição de treinos corresponde à fase em que os modelos de *Machine Learning* previamente treinados são integrados a uma interface computacional e utilizados para gerar predições a partir das informações fornecidas pelo usuário. Essa etapa tem como objetivo disponibilizar, de forma clara e acessível, as estimativas de tempo de corrida para uma distância definida pelo usuário, com base nos dados históricos utilizados no treinamento.

Após a etapa de coleta, processamento dos dados e treinamento, os modelos são posteriormente carregados pela aplicação no momento das predições. Dessa forma, o sistema é capaz de utilizar simultaneamente diferentes algoritmos de regressão previamente ajustados.

Para a geração das previsões, o usuário informa, na interface gráfica, a distância desejada para a corrida, conforme ilustrado na Figura 4.5. Esse valor é recebido pelo

backend e convertido para metros, de modo a manter compatibilidade com o formato dos dados utilizados durante o treinamento dos modelos. A partir disso, constrói-se uma estrutura de entrada padronizada, compatível com as variáveis do conjunto de treinamento, garantindo que todas as predições sejam realizadas de forma consistente entre os diferentes modelos.

Cada modelo carregado realiza, de forma independente, a predição do tempo estimado para percorrer a distância informada. As saídas dos modelos correspondem à duração prevista da atividade e são convertidas para o formato horas, minutos e segundos (h:min:s), facilitando a interpretação pelo usuário final. A ferramenta exibe os resultados gerados por todos os modelos de *Machine Learning* utilizados.

A apresentação simultânea das estimativas permite ao usuário comparar os resultados obtidos por diferentes abordagens de regressão, evidenciando possíveis variações entre os modelos. Essa estratégia contribui para maior transparência e reforça seu caráter de apoio à tomada de decisão, permitindo que corredores amadores tenham uma visão mais ampla das possíveis previsões de desempenho.

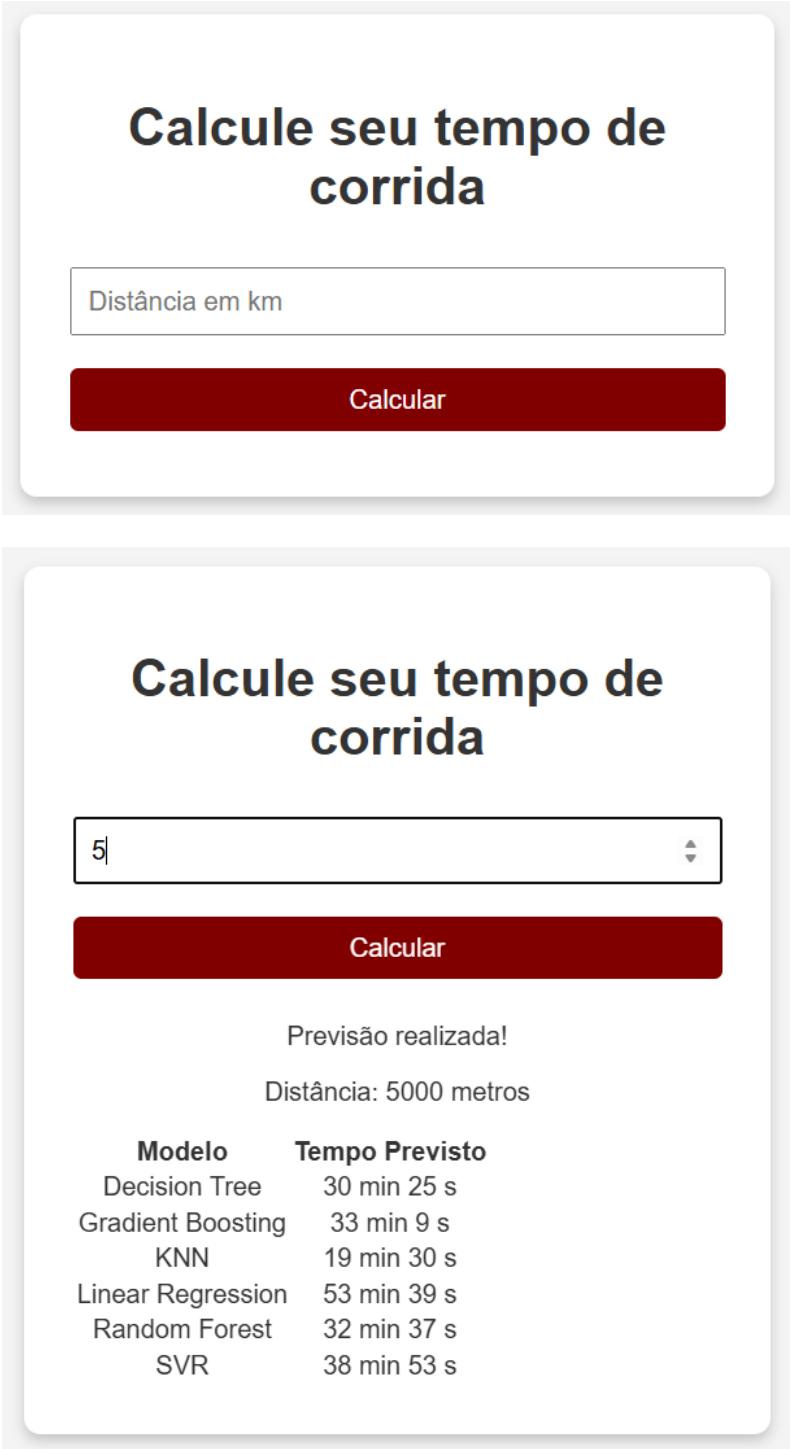

Figura 4.5: Interfaces para predição

5 Considerações Finais e Trabalhos Futuros

Este estudo teve como objetivo avaliar a aplicabilidade de modelos de *Machine Learning* na previsão do tempo de corrida a partir de dados provenientes de plataformas de monitoramento esportivo, especificamente do Samsung Health e do Strava. Os resultados obtidos demonstram que a combinação de algoritmos adequados e dados de qualidade é determinante para a obtenção de previsões precisas e fisiologicamente plausíveis.

De forma consistente, os modelos baseados em métodos de *ensemble*, especialmente o *Gradient Boosting*, apresentaram o melhor desempenho preditivo em ambas as bases de dados, com erros menores e maior capacidade de explicar a variabilidade do tempo de corrida. Esses resultados reforçam achados da literatura que apontam a superioridade de abordagens não lineares para modelar fenômenos complexos associados ao desempenho esportivo.

A comparação entre as plataformas evidenciou que os modelos treinados com dados do Strava apresentaram maior estabilidade, maior coerência fisiológica e maior capacidade de generalização, sobretudo para distâncias longas. Tal comportamento está diretamente relacionado à maior abrangência do conjunto de dados, que inclui corridas de até aproximadamente 43 km. Em contraste, os dados do Samsung Health, limitados a distâncias de até cerca de 21 km, mostraram-se adequados para previsões em distâncias curtas e intermediárias, mas insuficientes para extrapolações confiáveis em esforços prolongados.

A análise detalhada das estimativas de tempo ao longo de diferentes distâncias demonstrou que métricas tradicionais de avaliação, como MAE, RMSE e R^2, embora essenciais, não são suficientes para validar modelos no contexto esportivo. A coerência fisiológica das previsões e o respeito às relações esperadas entre distância e tempo mostraram-se critérios fundamentais para a avaliação da utilidade prática dos modelos.

Além disso, modelos como KNN e SVR apresentaram limitações significativas, produzindo estimativas inconsistentes e implausíveis, o que evidencia que nem todos os algoritmos amplamente utilizados em *Machine Learning* são adequados para problemas de

previsão de desempenho esportivo, especialmente quando há necessidade de extrapolação.

Apesar dos resultados promissores, este trabalho apresenta algumas limitações que devem ser consideradas. Primeiramente, observa-se uma limitação relacionada à homogeneidade dos dados, especialmente no conjunto do Samsung Health, no qual algumas variáveis foram preenchidas por médias ou apresentaram baixa variabilidade. Esse processo pode reduzir a capacidade dos modelos de capturar diferenças reais entre os treinos e limitar sua sensibilidade a fatores fisiológicos e ambientais que influenciam o desempenho.

Além disso, os modelos foram treinados a partir de dados históricos de um conjunto restrito de usuários, o que limita a generalização dos resultados para populações mais amplas e cenários distintos.

Por fim, a capacidade de extrapolação dos modelos é inerentemente limitada ao domínio dos dados de treinamento. Em particular, a ausência de corridas longas no conjunto do Samsung Health compromete a confiabilidade das previsões para distâncias maiores, enquanto mesmo o Strava, embora mais abrangente, não garante cobertura uniforme de todos os tipos de prova e perfis de corredores.

Em síntese, apesar das limitações encontradas, os resultados deste trabalho indicam que a previsão do tempo de corrida por meio de *Machine Learning* é viável e promissora, desde que sejam considerados cuidadosamente o tipo de modelo, a qualidade dos dados e os limites do domínio de treinamento.

Entre as perspectivas para trabalhos futuros, inicialmente, recomenda-se ampliar e diversificar o conjunto de dados, incluindo um maior número de atletas, diferentes níveis de desempenho e uma distribuição mais equilibrada de distâncias, especialmente para corridas longas, no caso do Samsung Health.

Outra possibilidade consiste na segmentação dos modelos por faixa de distância (por exemplo, curta, média e longa), reduzindo a necessidade de extrapolações e, potencialmente, aumentando a precisão das previsões. De forma complementar, uma maior exploração de variáveis fisiológicas, como a frequência cardíaca média, a variabilidade da frequência cardíaca e indicadores de recuperação, pode contribuir para uma modelagem mais robusta do desempenho.

Trabalhos futuros também podem explorar modelos temporais e sequenciais,

como redes neurais recorrentes ou modelos baseados em séries temporais, capazes de incorporar a evolução do treinamento ao longo do tempo e os efeitos cumulativos da carga. Adicionalmente, recomenda-se investigar o uso de algoritmos fundamentados em medidas de distância estatística, como abordagens do tipo SafeML[4], com o objetivo de quantificar a incerteza das previsões, por exemplo por meio da estimativa do desvio padrão das respostas dos modelos de *Machine Learning* em relação à distribuição dos dados de treinamento. Esse tipo de análise pode fornecer indicadores de confiabilidade das previsões, contribuindo para decisões mais seguras em cenários de extrapolação.

Outra possibilidade consiste em apresentar um modelo conceitual entidade- relacionamento que descreva as estruturas de dados manipuladas pela aplicação desenvolvida, oferecendo uma visão clara da organização, relacionamento e integridade dos dados, o que pode apoiar futuras evoluções do sistema.

Além disso, técnicas avançadas de validação, como a validação cruzada estratificada por distância ou por atleta, podem fornecer avaliações mais realistas da capacidade de generalização dos modelos.

Por fim, a aplicação prática dos modelos em sistemas de treino de recomendação ou em ferramentas de apoio à tomada de decisão para atletas e treinadores representa um caminho promissor, ampliando o impacto prático da pesquisa e contribuindo para o desenvolvimento de soluções inteligentes no contexto do esporte e da saúde.

[4]O termo "SafeML" é usado aqui de forma genérica para se referir a abordagens que estimam a confiabilidade das previsões com base na distância estatística dos dados em relação ao conjunto de treinamento, sem necessariamente se referir a uma implementação específica.

Bibliografia

AL., M. et. Velocidade pico ou crítica determinadas em campo: predizendo performance em corredores recreacionais. 2021. Frontiers in Physiology summary.

AL., P. et. Perfil e prevalência de lesões em corredores amadores de são luís-ma. *RBPFEX*, 2025.

AL., V. et. How do runners experience personalization of their training scheme: The inspirun e-coach? *Sensors*, 2020.

AL., V. et. Personalization of training sessions in running apps: Insights from pmc. *PMC Article*, 2020.

ANTONIO, S.; LAUX, C. Educação física: Ciência e perspectiva — definição de corredores amadores e recreacionais. 2022. Documento institucional.

BACA, A.; KORNFEIND, P. *Technology in Sports: Advanced Monitoring and Training Systems*. [S.l.]: Springer, 2011.

BEIS, L. Y. et al. Environmental conditions and marathon performance: Machine learning analysis of 668,509 runners in the berlin marathon (1999–2019). *Med Sci Sports Exerc*, v. 55, n. 7, p. 1234–1245, 2023.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, v. 13, p. 281–305, 2012.

BREIMAN, L. Random forests. *Machine Learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L. et al. *Classification and Regression Trees*. [S.l.]: Wadsworth, 1984.

BUCHHEIT, M.; LAURSEN, P. B. Monitoring training status with heart rate measures: do all roads lead to rome? *Frontiers in Physiology*, 2013.

CHAI, T.; DRAXLER, R. Root mean square error (rmse) or mean absolute error (mae)? *Geoscientific Model Development*, v. 7, p. 1247–1250, 2014.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, 1967.

DAVIDSON, P. et al. Smartwatch-derived data and machine learning algorithms estimate classes of ratings of perceived exertion in runners: A pilot study. *Sensors*, v. 20, n. 9, p. 2637, 2020. Disponível em: ⟨https://www.mdpi.com/1424-8220/20/9/2637?⟩

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. [S.l.]: Springer, 2009.

HONG, D.; SUN, S. Machine learning regressors to estimate continuous oxygen uptakes (vo). *Applied Sciences*, v. 14, n. 17, p. 7888, 2024. Disponível em: ⟨https://www.mdpi.com/2076-3417/14/17/7888⟩.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, v. 22, n. 4, p. 679–688, 2006.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*. [S.l.: s.n.], 1995.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill, 1997.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Applied Linear Regression Models*. [S.l.]: Wiley, 2012.

NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika*, v. 78, n. 3, p. 691–692, 1991.

NETTO, C. G. Pesquisa analisa o perfil de corredores de rua amadores. In: . [S.l.: s.n.], 2017. Unicamp news.

NUGROHO, A. A. Intelligence sports performance analytics from strava using big data platform with multi-layer perceptron. *Engineering Headway*, v. 27, p. 364–379, 2025. Disponível em: ⟨https://www.scientific.net/EH.27.364⟩.

OH, H. et al. What is ehealth? a systematic review of published definitions. *Journal of Medical Internet Research*, v. 7, p. e1, 2005.

OLIVEIRA, C.; SANTOS, A. Aplicação de técnicas de aprendizagem de máquina para apoio a atletas amadores. In: *Anais do Congresso Brasileiro de Computação Aplicada ao Esporte*. [S.l.]: SBC, 2022. p. 55–67.

Open-Meteo. *Open-Meteo: Free Open-Source Weather API*. 2024. Acesso em: jan. 2026. Disponível em: ⟨https://open-meteo.com/⟩.

PATEL, M. S.; ASCH, D. A.; VOLPP, K. G. Wearable devices as facilitators, not drivers, of health behavior change. *Journal of the American Medical Association*, v. 313, p. 459–460, 2012.

PIRSCOVEANU, C.-I.; OLIVEIRA, A. S. Prediction of instantaneous perceived effort during outdoor running using accelerometry and machine learning. *European Journal of Applied Physiology*, v. 124, p. 963–973, 2023.

REIS, P. H. A. C. Monografia (Graduação em Ciência da Computação), *Uso de Machine Learning no Esporte: Apoio Inteligente para Corredores não Profissionais*. Juiz de Fora: [s.n.], 2025.

SMITH, J.; OTHERS. Training volume and training frequency changes associated with boston marathon race performance. *Sports Medicine*, v. 55, p. 101–115, 2025. Disponível em: ⟨https://link.springer.com/article/10.1007/s40279-025-02304-4⟩.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. *Statistics and Computing*, v. 14, p. 199–222, 2004.

STRAVA. Relatório anual strava 2024: Corrida é o esporte mais praticado no mundo. *Strava Insights*, 2024. Disponível em: ⟨https://blog.strava.com/pt/⟩ Acesso em: 24 dez. 2025.

STUDY, A. of P. Study of the motivation of spanish amateur runners based on training patterns and gender. *Journal Article*, 2020. PubMed.

SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. [S.l.]: MIT Press, 2018.

TOPOL, E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. [S.l.]: Basic Books, 2019.

WANG, Y.; XU, D. User engagement with digital health wearables in sport and fitness. *Sport Sciences Review*, v. 29, p. 101–120, 2020.

WEISZ, N. et al. Machine learning predicts recovery in endurance athletes but requires personalized strategies. *European Journal of Applied Physiology*, v. 124, p. 1–15, 2024. Disponível em: ⟨https://link.springer.com/article/10.1007/s00421-024-05530-2⟩.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error over the root mean square error. *Climate Research*, v. 30, p. 79–82, 2005.