Uso de Machine Learning no Esporte: Apoio Inteligente para Corredores não Profissionais

Pedro Henrique Almeida Cardoso Reis

JUIZ DE FORA

MARÇO, 2025

Uso de Machine Learning no Esporte: Apoio Inteligente para Corredores não Profissionais

Pedro Henrique Almeida Cardoso Reis

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Victor Ströele de Andrade Menezes

JUIZ DE FORA

MARÇO, 2025

Uso de Machine Learning no Esporte: Apoio Inteligente para Corredores não Profissionais

Pedro Henrique Almeida Cardoso Reis

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:

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

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

Regina Maria Maciel Braga
Doutora em Engenharia de Sistemas e Computação (UFRJ)

JUIZ DE FORA
10 DE MARÇO, 2025

Aos meus amigos e irmãos.

Aos pais, pelo apoio e sustento.

Resumo

Este trabalho de conclusão de curso aborda a aplicação de técnicas de aprendizado de máquina no esporte, com foco em corredores amadores e entusiastas. O objetivo principal é desenvolver um sistema inteligente de apoio para auxiliar corredores na melhoria de seu desempenho durante a prática da corrida. Através da análise de dados coletados de *smartwatches*, o sistema oferece *insights* personalizados, prevendo o tempo estimado para percorrer uma distância específica. Com base nesses *insights*, o atleta pode ajustar seu ritmo de corrida, otimizando sua performance de maneira mais eficaz. Com isso, o trabalho busca beneficiar corredores não profissionais, ajudando-os a otimizar seu rendimento e a alcançar melhores resultados em suas corridas.

Palavras-chave: Aprendizado de Máquina, ehealth, corredores não profissionais, temperatura, qualidade do sono

Abstract

This final project addresses the application of machine learning techniques in sports, focusing on amateur and enthusiastic runners. The main objective is to develop an intelligent support system to help runners improve their performance, prevent injuries and optimize their training. By analyzing data collected from smartwatches, the system offers personalized insights, predicting the estimated time to cover a specific distance. Based on these insights, the athlete can adjust their running pace, optimizing their performance more effectively. With this, the work seeks to benefit non-professional runners, helping them optimize their performance and achieve better results in their races.

Keywords: machine learning, e-health, non-professional runners, data analysis, weather, sleep quality

Agradecimentos

Primeiramente, gostaria de agradecer à Deus, meus pais, minha irmã, minha avó e meu falecido avô pelo apoio e sustento durante esses longos anos fora de casa. Sem essa base, nada disso seria possível.

Aos meus amigos que fiz durante todo esse percurso, obrigado pela ajuda nos momentos de dificuldade. Sempre estiveram presente nos momentos difíceis mas também nos momentos de descontração.

Em seguida, quero estender minha gratidão ao meu orietador Victor Ströele, pela paciência e por todos os ensinamentos. Seu apoio foi fundamental para a conclusão deste trabalho.

E, por fim, um agradecimento especial para todos os professores do Departamento de Ciência da Computação da Universidade Federal de Juiz de Fora por todo o ensinamento passado nesses longos anos de estudo.

"A união entre esporte e tecnologia não é o futuro; é o presente em constante evolução.".

Inspirado em Sebastian Coe

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

1 Introdução

1.1 Apresentação do tema

Segundo (LOTTENBERG, 2021), após a pandemia de Covid-19, uma transformação significativa nos hábitos relacionados à saúde tem ocorrido não só entre os brasileiros mas também no mundo intero. A crise sanitária global provocada pela Covid-19 instigou profundas reflexões sobre a saúde física e, em especial, ressaltou a importância do autocuidado (TENORIO, 2020). Esta pandemia emergiu como um dos principais catalisadores na promoção de uma consciência mais aguçada acerca da necessidade de cuidar do bem-estar pessoal de forma mais diligente.

Essa crescente conscientização abriu caminho para um interesse ampliado em hábitos saudáveis, tais como uma alimentação equilibrada, a garantia de um sono adequado e a incorporação regular de exercícios físicos. Diante dessa nova realidade, é patente que não podemos mais retomar o mesmo paradigma que prevalecia em nossa abordagem à saúde antes de março de 2020.

Pesquisas recentes têm revelado um aumento notável na preocupação dos brasileiros com a saúde física (VERDE, 2023). Dentro deste contexto, a prática da corrida tem despontado como uma das atividades esportivas mais procuradas. Isso ocorre em virtude dos inúmeros benefícios associados a essa modalidade esportiva, que incluem a prevenção de problemas cardiovasculares, o combate à depressão e o aprimoramento da capacidade de memória (BRUCE, 2023). Nesse cenário, é evidente que a corrida não é apenas uma atividade física em ascensão, mas também uma tendência que está diretamente ligada à crescente preocupação dos brasileiros com seu bem-estar físico.

Ao analisarmos as abordagens eficazes no contexto abordado, percebemos uma crescente tendência em adotar estratégias que envolvem o desenvolvimento de ferramentas de apoio personalizadas para indivíduos. Estas ferramentas desempenham um papel crucial ao coletar informações relevantes, as quais, por sua vez, contribuem para otimizar o desempenho durante o treinamento esportivo e identificar possíveis causas de variações

na performance durante a prática esportiva. Geralmente, referidas como aplicações de e-health (EYSENBACH, 2001), essas ferramentas assumem a forma de sistemas que monitoram continuamente o usuário, utilizando os dados coletados para fornecer respostas esclarecedoras e significativas.

Essas soluções tecnológicas representam uma abordagem inovadora que demonstram grande potencial para melhorar o desempenho esportivo e a saúde dos praticantes, abrindo novas perspectivas para a pesquisa e desenvolvimento no campo de e-health aplicada ao esporte.

No entanto, apesar do imenso potencial apresentado pelas tecnologias de e-health, é essencial destacar que a implementação dessas aplicações exige a observância de critérios de qualidade rigorosos, conforme estabelecido pela norma internacional (International Organization for Standardization, 2011). É de fundamental importância que tais aplicações alcancem um alto padrão de qualidade em termos de disponibilidade de serviço, eficiência no processamento dos dados coletados e segurança no acesso às informações dos usuários, como também estabelecido na (International Organization for Standardization, 2011). Além disso, essas ferramentas e-health necessitam de uma abordagem multidisciplinar, envolvendo profissionais de saúde, tecnologia da informação e segurança da informação, a fim de garantir que as aplicações no contexto esportivo sejam seguras, eficazes e capazes de fornecer resultados confiáveis conforme a própria ONU discute em (World Health Organization, 2005).

Portanto, o desenvolvimento de aplicações neste domínio específico requer uma abordagem criteriosa, que inclui a consideração de múltiplas medidas para garantir que os recursos de apoio estejam prontamente disponíveis para os usuários.

1.2 Contextualização

A prática de corrida e a busca por um estilo de vida mais saudável têm se tornado cada vez mais populares, especialmente entre pessoas não profissionais que buscam melhorar sua condição física e bem-estar (BONAT, 2024). Atualmente existem diversas aplicações e-health como o *Strava, Samsung Health, adidas Running* responsáveis por coletar métricas de corrida como distância, tempo, velocidade e frequência cardíaca. Essas aplicações,

em sua grande maioria, utilizam-se de uma combinação de sensores para coletar os mais diferenciados dados de treino do usuário. Após a coleta dos dados, é retornado para o atleta todas as métricas obtidas durante o treino.

No âmbito deste estudo, é utilizada uma ampla variedade de algoritmos de Machine Learning com o objetivo de prever o tempo que um atleta amador levará para percorrer determinada distância. Para otimizar a acurácia das predições, foi realizada uma análise detalhada dos dados de treinamento coletados durante a prática de corrida. A escolha e aplicação desses algoritmos não apenas aprimoram a precisão dos resultados, mas também fornecem *insights* valiosos para atletas amadores.

Além do mais, este estudo representa uma contribuição para a interseção entre tecnologia e saúde. À medida que mais pessoas adotam a corrida como parte integrante de seu estilo de vida, o aumento da eficácia do treinamento se tornam fatores cruciais para manter a continuidade desse hábito. Futuros trabalhos podem expandir essa abordagem, incorporando novas tecnologias e métodos de análise, a fim de promover não apenas a performance esportiva, mas também o bem-estar e a qualidade de vida dos praticantes.

1.3 Descrição do problema

Corredores não profissionais muitas vezes enfrentam desafios relacionados ao planejamento e otimização de seus treinos, bem como à prevenção de lesões. A falta de orientação personalizada e o excesso de informações provenientes de aplicativos de rastreamento e dispositivos de monitoramento podem sobrecarregar os corredores, dificultando o alcance de seus objetivos de forma segura e eficaz.

Foi observado também que a individualidade de cada corredor, incluindo seu nível de condicionamento físico, objetivos e histórico de lesões, muitas vezes não é adequadamente considerada nas abordagens tradicionais de treinamento. Isso levanta a necessidade de um sistema inteligente de apoio aos corredores não profissionais que possa utilizar técnicas de Machine Learning para analisar e interpretar os dados coletados durante os treinos, proporcionando orientações personalizadas que otimizem o desempenho, minimizem o risco de lesões e tornem a experiência da corrida mais gratificante e eficaz.

O contexto atual de conectividade digital proporciona uma oportunidade única

para a implementação de soluções baseadas em Machine Learning. A crescente disponibilidade de dados relacionados à atividade física, como informações de rastreamento de GPS, batimentos cardíacos, padrões de sono, nutrição e temperatura, abre portas para o desenvolvimento de sistemas inteligentes que podem fornecer *insights* valiosos. Essas soluções podem considerar não apenas o desempenho durante os treinos, mas também fatores que afetam a recuperação e a manutenção da saúde a longo prazo.

1.4 Justificativa/Motivação

A demanda por estratégias de treinamento personalizadas e eficazes tem crescido em um mundo cada vez mais voltado para a saúde e o bem-estar. Nesse contexto, o uso de Machine Learning como uma ferramenta de apoio inteligente apresenta uma solução promissora para oferecer um suporte para corredores não profissionais, visando otimizar seu desempenho durante as corridas.

Os avanços recentes, especialmente no campo da Internet das Coisas, têm proporcionado um ambiente propício para o desenvolvimento de soluções inovadoras no esporte. Os sensores de monitoramento em tempo real, que se tornaram acessíveis e eficientes, permitem a coleta de uma riqueza de dados durante as sessões de corrida. Com a adição de dispositivos de computação de borda, a capacidade de processamento desses dados em tempo real tornou-se mais acessível, uma vez que estamos processando os dados fisicamente mais perto de sua fonte, abrindo caminho para sistemas distribuídos e em funcionamento contínuo (YEUNG, 2023). Essas inovações tecnológicas oferecem uma oportunidade única para a criação de sistemas de apoio inteligentes que podem analisar o desempenho do corredor, identificar áreas de melhoria e fornecer orientações personalizadas para o treinamento, tornando o uso de Machine Learning no esporte uma abordagem altamente relevante e promissora.

1.5 Objetivos

O objetivo deste trabalho é investigar e demonstrar como o uso de técnicas de Machine Learning pode ser aplicado de forma inteligente e eficaz para oferecer suporte, mais espe-

cificamente, visando otimizar o desempenho de corredores não profissionais.

1.5.1 Objetivos Gerais

Temos como objetivo geral analisar e desenvolver um sistema baseado em técnicas de Machine Learning para oferecer suporte inteligente a corredores não profissionais, visando aprimorar seu desempenho. Isso foi realizado por meio da criação e análise de modelos de Machine Learning que utilizam dados relevantes, como histórico de treinamento, métricas de desempenho e características individuais dos corredores, contribuindo assim para o aprimoramento da performance esportiva e o bem-estar dos corredores não profissionais.

1.5.2 Objetivos Específicos

Temos como objetivo específico deste trabalho:

- Realizar um levantamento bibliográfico para compreender as principais técnicas de Machine Learning aplicáveis ao monitoramento e apoio a atletas amadores na prática da corrida.
- Coletar e analisar dados relevantes de corredores não profissionais, como frequência cardíaca, passadas, distância percorrida, entre outros, a fim de identificar padrões e *insights* que possam influenciar o desempenho e a saúde do corredor.
- Avaliar a eficácia do sistema proposto por meio de testes com corredores não profissionais.
- Documentar e apresentar os resultados obtidos, evidenciando a viabilidade, eficácia e potenciais benefícios do uso de técnicas de Machine Learning como suporte inteligente para corredores não profissionais, contribuindo para a melhoria da prática esportiva.

1.6 Organização do trabalho

Diante da introdução apresentada, este trabalho de conclusão de curso está organizado da seguinte forma: No **Capítulo 2** é apresentada a fundamentação teórica, com os conceitos

essenciais para a compreensão do estudo. No **Capítulo 3** é realizada a revisão bibliográfica, abordando as referências e estudos relacionados a este trabalho de conclusão de curso. No **Capítulo 4** é detalhado o processo de desenvolvimento do trabalho, incluindo as análises realizadas e os resultados obtidos. Por fim, o **Capítulo 5** apresenta as conclusões do estudo, bem como sugestões para possíveis trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo são apresentados os fundamentos para a compreensão dos tópicos abordados neste trabalho. Inicialmente, falaremos sobre corredores não profissionais e seus desafios no esporte. Discutiremos também os conceitos de *e-health*, *Computação Ubíqua* e *Machine Learning*.

Além disso, exploraremos a interseção desses conceitos e como eles se relacionam com os objetivos e desafios deste estudo. Vamos examinar como a integração de *e-health* com a computação ubíqua e técnicas de Machine Learning está impactando a prestação de serviços de saúde, melhorando o monitoramento de pacientes e permitindo avanços significativos na área médica.

A aplicação do conceito de *e-health* no contexto de corredores amadores envolve o uso de tecnologias para monitoramento e análise de dados de corrida, provenientes de smartwatches. A *Computação Ubíqua* é fundamental para o desenvolvimento de soluções que integrem dispositivos inteligentes de forma contínua e transparente, permitindo a coleta de dados em tempo real durante os treinos. Por fim, o uso de Machine Learning possibilita a análise desses dados, com foco na previsão de desempenho. Dessa forma, os três conceitos são abordados de maneira integrada para fornecer um sistema inteligente que apoia corredores amadores visando otimizar seu desempenho

2.1 Atletas amadores

A utilização de Machine Learning no esporte tornou-se, nos últimos anos, uma área de pesquisa de grande relevância. Uma aplicação promissora para essa pesquisa é a utilização de Machine Learning voltada para corredores amadores. Segundo a OMS, entende-se por atletas amadores o indivíduo que pratica alguma atividade física pelo menos 3 vezes por semana, com duração mínima de 30 minutos (PIEDRAS, 2021). Essa definição ampla abrange uma variedade de corredores, desde aqueles que estão começando até aqueles que já têm alguma experiência no esporte.

Diante desse cenário, essa pesquisa surge com o objetivo de auxiliar corredores não profissionais a prevenir lesões, oferecer treinamento personalizado e entender situações no qual o treinamento não rendeu como o esperado. Para fundamentar ainda mais essa pesquisa, é relevante explorar exemplos de estudos anteriores que aplicaram Machine Learning no esporte, em especial na corrida.

Em (KNECHTLE BEAT; DI GANGI, 2019) é feito um estudo comparando como as condições climáticas afetaram a performance dos atletas na maratona de *Boston*. Nesse artigo, é demonstrado como variáveis externas, como o clima, podem influenciar o desempenho esportivo. Nesse âmbito, técnicas de Machine Learning podem ser aplicadas para entender e mitigar esses efeitos.

2.2 e-health

Como proposto por (EYSENBACH, 2001), e-health é um conceito amplo. Pode-se dizer que e-health é um termo que se refere ao uso da tecnologia da informação e comunicação, especialmente a internet e sistemas eletrônicos, para melhorar a prestação de serviços de saúde, o gerenciamento de informações de saúde e a entrega de cuidados médicos.

Atualmente, observamos um aumento significativo na utilização de sistemas de monitoramento e suporte à saúde das pessoas, impulsionado pelo avanço tecnológico. Dispositivos vestíveis, como relógios inteligentes e medidores de saúde conectados à internet, estão se tornando mais acessíveis e difundidos, permitindo o acompanhamento contínuo de parâmetros de saúde, como frequência cardíaca e atividade física.

Com base no que foi exposto acima, é fundamental destacar que este estudo tem como objetivo proporcionar *insights* valiosos para corredores não profissionais, visando capacitá-los a antecipar o desempenho de seus próximos treinos e evitar lesões. Nesse contexto, é de extrema importância a colaboração de um profissional de educação física capaz de orientar o atleta com base nas informações obtidas.

2.3 Computação Ubíqua

Em (WEISER, 1991) o termo Computação Ubíqua foi cunhado para descrever a ideia de dispositivos interconectados em todo lugar, de maneira tão imperceptível para os seres humanos que, eventualmente, deixaríamos de perceber sua presença. Nesse contexto, os dispositivos, sensores e sistemas de computação estão presentes em todos os lugares e fazem parte integrante da vida cotidiana das pessoas, sem que elas necessariamente percebam sua presença.

Ao explorar a área de serviços de *e-health*, naturalmente, podemos visualizar um cenário em que uma variedade de sensores coleta diversos sinais do paciente. Paralelamente a essa coleta, ocorrem diversos processos de computação simultaneamente, com o objetivo de disponibilizar análises relevantes. É justamente esse arranjo de processos e dispositivos que estão sendo categorizado como Computação Ubíqua. Isso implica que, no contexto de *e-health*, a tecnologia se torna parte integrante da vida do atleta, coletando dados de treino de forma discreta e realizando análises em segundo plano para melhorar a performance e oferecer *insights* valiosos.

O monitoramento de atividade física representa um campo promissor para a aplicação de sistemas de computação ubíqua, contudo, algumas questões demandam uma atenção minuciosa. Primeiramente, destaca-se a importância da proteção da privacidade dos dados dos usuários, tornando-se essencial assegurar a segurança e a confidencialidade dessas informações. Além disso, a dependência de sensores e dispositivos interconectados na computação ubíqua é um fator crítico. A confiabilidade desses dispositivos é de extrema relevância, uma vez que falhas técnicas podem impactar negativamente a experiência do corredor, comprometendo os benefícios pretendidos. Portanto, ao abordar essas questões de privacidade e confiabilidade, é possível criar um ambiente seguro e eficaz para a aplicação bem-sucedida da computação ubíqua no monitoramento esportivo.

2.4 Machine Learning

Machine Learning é um subcampo da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos estatísticos que permitem aos computadores aprender e

melhorar o desempenho em tarefas específicas a partir de dados, ou seja, sua ideia central é a de que o algoritmo consiga, por meio de tentativa e erro, se tornar cada vez melhor em cumprir a tarefa desejada.

Os algoritmos de Machine Learning podem ser categorizados em três principais tipos: Aprendizado Supervisionado, Aprendizado Não Supervisionado e Aprendizado por Reforço. No **Aprendizado Supervisionado** o modelo aprende a partir de um conjunto de dados rotulado, onde cada entrada possui uma saída conhecida. Ele busca identificar padrões e fazer previsões com base nos exemplos fornecidos. Temos como principais exemplos os algoritmos de *Regressão Linear, Regressão Logística, Máquinas de Vetores de Suporte (SVM), Árvores de Decisão, Florestas Aleatórias e Redes Neurais Artificiais.*

Já no **Aprendizado Não Supervisionado** o modelo recebe dados sem rótulos e precisa encontrar padrões por conta própria, agrupando informações semelhantes ou reduzindo a dimensionalidade dos dados, tendo como principais exemplos *Algoritmo K-Means (Clusterização) e Análise de Componentes Principais (PCA).*

Por fim, temos o **Aprendizado por Reforço**, one o algoritmo aprende por meio de tentativa e erro, recebendo recompensas ou penalidades para aprimorar suas decisões ao longo do tempo. Temos como exemplo *Q-Learning e Deep Q-Networks (DQN).*

Neste trabalho fizemos uso de modelos preditivos (Aprendizado Supervisionado). Esses modelos são usados para prever valores ou categorias com base em padrões extraídos de dados históricos. Eles podem ser divididos em dois tipos principais: *Modelos de Regressão* e *Modelos de Classificação.* Neste trabalho foi usando um modelo preditivo de regressão para estimar o tempo que o corredor levaria para percorrer uma determinada distância com base em variáveis como distância, temperatura e tempo de sono. Desse modo, o modelo analisa padrões nos dados históricos do corredor e projeta um tempo estimado para um novo percurso.

Em (ALSAREII et al., 2022) são feitas comparações com diferentes algoritmos de Machine Learning para monitoramento e classificação de atividades físicas. Os resultados obtidos fornecem *insights* valiosos para o desenvolvimento de um sistema de apoio inteligente voltado para corredores não profissionais.

2.4.1 overfitting

O overfitting é um fenômeno comum em modelos de aprendizado de máquina, incluindo Árvores de Decisão (Decision Trees), e ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, capturando não apenas os padrões subjacentes, mas também o ruído e as flutuações aleatórias presentes nesses dados. Como consequência, o modelo perde a capacidade de generalizar para novos dados, resultando em um desempenho inferior em conjuntos de teste ou em situações reais.

No contexto das Árvores de Decisão, o overfitting tende a acontecer quando a árvore cresce muito profundamente, criando ramificações excessivas e regras muito específicas para os dados de treinamento. Isso pode levar a uma estrutura complexa que se adapta perfeitamente aos dados de treinamento, mas falha ao lidar com dados não vistos anteriormente. Por exemplo, uma árvore muito profunda pode criar regras que se aplicam a casos extremamente específicos, como outliers ou erros de medição, o que prejudica sua capacidade de prever corretamente novos exemplos.

2.5 Transferência de Aprendizado

A *Transferência de Aprendizado (Transfer Learning)* é uma técnica de *Machine Learning* em que um modelo treinado em um conjunto de dados e tarefa específicos é reutilizado, total ou parcialmente, para um novo problema. Essa abordagem reduz a necessidade de treinamento extensivo do modelo a partir do zero, economizando tempo computacional e melhorando a eficiência do aprendizado, especialmente quando os dados disponíveis são limitados.

Modelos pré-treinados em grandes conjuntos de dados esportivos podem ser ajustados para prever tempos de corrida, através das métricas de corrida extraída dos smartwatches de outros usuários. Essa abordagem permite que o sistema utilize conhecimento adquirido de atletas ou de bases de dados maiores para fornecer previsões mais precisas e personalizadas, mesmo quando há uma quantidade limitada de dados disponíveis para um corredor específico.

3 Revisão Bibliográfica

Em (ALSAREII et al., 2022) é feita a comparação de várias técnicas de aprendizado de máquina para identificar a mais apropriada na classificação de atividades em um conjunto de dados de atividade física intencionalmente desequilibrado. Ele avalia seis classificadores bem conhecidos e analisa o desempenho de algoritmos de aprendizagem com diferentes divisões de treinamento e graus de desequilíbrio para identificar as técnicas de aprendizado de máquina mais adequadas.

Neste artigo temos como ponto forte o uso de algoritmos de Machine Learning para a classificação de diferentes atividades: *sentado, em pé, andando, deitado, subindo escadas e descendo escadas.* Os algoritmos como um todo conseguiram uma boa eficácia, entretanto, não foi abordado atividades físicas mais intensas, como a corrida.

Já em (KNECHTLE BEAT; DI GANGI, 2019) é apresentada uma discussão conceitual sobre um atleta que corre. Existem hoje muitas pessoas que em grande parte do tempo treinam para corridas em academias climatizadas, academias essas que proporcionam um menor desgaste do corpo. Este estudo examinou a relação entre as condições meteorológicas, juntamente com o sexo e o país de origem, com o desempenho na Maratona de Boston de 1972 a 2018. A análise dos mais de 300.000 participantes foi feita usando *Generalized Additive Mixed Models*, uma extensão dos Modelos Lineares Generalizados Aditivos que incorporam a capacidade de modelar relações não lineares entre variáveis independentes e dependentes. Sendo que somente os atletas que tiveram os melhores tempos foram analisados. Este estudo tem como ponto forte os fatores climáticos que afetam de forma negativa e positiva o desempenho de atletas. Todavia, o estudo avaliou somente a performance dos melhores colocados, aqueles atletas vencedores, e aqueles que chegaram próximo a eles, deixando de lado os atletas amadores.

Em (MANTZIOS et al., 2021) é apresentada uma abordagem semelhante a apresentada no trabalho anterior. Foram analisadas 1.258 corridas realizadas entre 1936 e 2019 em 84 localidades e 42 países diferentes. O artigo se concentra no desempenho de atletas em corridas de resistência, que geralmente envolvem distâncias longas, como ma-

ratonas, ultra-maratonas, ou corridas de longa distância em trilhas. Este estudo avaliou como parâmetros climáticos individuais, ou combinados, (temperatura, umidade, velocidade do vento e carga solar) afetaram o desempenho máximo durante eventos de corrida de resistência. Nesse estudo, foram coletados dados de maratonas, marcha atlética de 50 km, marcha atlética de 20 km e provas de 10.000, 5.000 e 3.000m . As Árvores de decisão de Machine Learning mostrou que a temperatura do ar foi o parâmetro climático mais importante. A conclusão obtida é a mesma do artigo anterior: temperatura, umidade, vento, podem influenciar o desempenho de um atleta durante a corrida.

O estudo conduzido por Ciabattoni et al. (2017) apresenta uma abordagem inovadora para a detecção de estresse mental em voluntários por meio da utilização de *smartwatches*. A pesquisa parte do pressuposto de que o estresse é um fator significativo que impacta tanto a saúde individual quanto o desempenho em diversas atividades do dia a dia. Dessa forma, a capacidade de monitorar o estresse de maneira contínua e não invasiva pode trazer benefícios substanciais para diferentes áreas, como saúde, produtividade e até mesmo o desempenho esportivo.

Inicialmente, os autores realizam uma ampla revisão da literatura sobre o estresse e seus impactos fisiológicos, além de abordarem estudos anteriores que exploram a detecção desse estado emocional por meio de sensores biométricos. Nesse contexto, os *smartwatches* surgem como uma alternativa promissora devido à sua capacidade de coletar dados fisiológicos em tempo real, de maneira conveniente e acessível. A proposta do estudo consistiu em submeter os voluntários a tarefas cognitivas que demandassem diferentes níveis de esforço mental, enquanto os dados capturados pelos *smartwatches* eram registrados e analisados. Posteriormente, essas informações foram utilizadas para treinar modelos de Machine Learning com o objetivo de prever se um voluntário estava ou não em um estado de estresse.

Esse estudo é particularmente relevante para a pesquisa em andamento, pois destaca a confiabilidade dos *smartwatches* na captura de dados fisiológicos e reforça sua aplicabilidade para o monitoramento de métricas esportivas. No contexto da corrida, métricas como frequência cardíaca, velocidade e gasto calórico são essenciais para avaliar o desempenho do atleta.

Em (ALSAREII et al., 2022), os autores descrevem um processo de computação de recursos a partir de sinais brutos de acelerômetro 3D e giroscópio obtidos de *smartphones*. Esses recursos são usados em um algoritmo de aprendizado de máquina para classificar Atividades da Vida Diária. As etapas de pré-processamento incluem filtragem passa-baixa, filtragem mediana e separação do sinal de aceleração em aceleração do corpo e aceleração gravitacional.

Além disso, o artigo também descreve experimentos que avaliam o desempenho de diferentes algoritmos de aprendizado de máquina em conjuntos de dados desequilibrados e equilibrados. O desequilíbrio de classes é uma preocupação importante no aprendizado de máquina, e diferentes algoritmos são testados para ver como eles lidam com essa questão. Os experimentos usam uma variedade de classificadores, incluindo *SVM, Gradient Boosting, Extreme Gradient Boostin g, CatBoost, AdaBoost com árvore de decisão e AdaBoost com Random Forest.*

Os resultados são avaliados usando a pontuação F-macro média como métrica de desempenho. O artigo fornece detalhes sobre os parâmetros usados para treinar esses classificadores e apresenta equações para calcular a pontuação F.

Após a análise dos trabalhos citados anteriormente, observa-se que os estudos enfatizam a importância do uso de Machine Learning para a interpretação dos dados coletados, demonstrando como modelos treinados adequadamente podem identificar padrões e fornecer *insights* personalizados para os usuários. Essa abordagem está diretamente alinhada com o objetivo da pesquisa em curso, que busca utilizar dados de *smartwatches* para prever o tempo necessário para percorrer determinadas distâncias, oferecendo assim recomendações personalizadas para corredores não profissionais.

4 Materiais e Métodos

Neste capítulo apresentamos detalhadamente o processo completo utilizado para conceber, desenvolver e implementar [1] um modelo de Machine Learning destinado a auxiliar corredores amadores. O objetivo principal é fornecer uma visão abrangente das etapas envolvidas no desenvolvimento deste sistema, desde a coleta de dados até a construção do modelo preditivo e sua avaliação.

A Figura 4.1 ilustra as etapas envolvidas no desenvolvimento do trabalho. Na etapa de **Coleta de Dados** o atleta captura suas métricas de corrida por meio de um *smartwatch*, e, posteriormente, os dados são exportados para análise. Na etapa de **Processamento dos Dados**, são realizadas a limpeza e a normalização dos dados, além do treinamento dos modelos de Machine Learning. Por fim, na etapa de **Apresentação dos Resultados**, os dados processados e as análises são disponibilizados ao usuário. Nessa etapa final auxiliamos corredores na melhoria de seu desempenho. Nas seções a seguir, cada uma dessas etapas é detalhada, explicando como foram desenvolvidas e implementadas.

Figura 4.1: Fluxograma das etapas do trabalho

A **coleta de dados** para este estudo foi realizada de maneira cuidadosa e estruturada, visando capturar informações relevantes para o desenvolvimento do modelo de apoio aos corredores não profissionais.

Foram coletados dados de um *smartwatch Galaxy Watch 4 Classic* de um participante voluntário. Esse dispositivo foi escolhido por sua capacidade de capturar informações em tempo real durante a atividade física, incluindo dados sobre distância per-

[1] https://github.com/PHenriqueCEC/PedroHenriqueAlmeidaTCC

corrida, frequência cardíaca, velocidade, altitude e outros parâmetros relevantes para a prática da corrida.

Além disso, aplicativos de corrida amplamente utilizados, mais especificamente o *Samsung Health*, foi empregado para exportar os dados dos usuários. O *Samsung health* forneceu informações valiosas sobre as rotas percorridas, a frequência cardíaca durante os treinos de corrida, temperatura ambiente durante a prática da corrida e dados de sono.

De uma maneira geral, o usuário exporta seus dados de corrida do *Samsung Health* e, a partir do arquivo compactado que é gerado, ele faz o *upload* desse arquivo na ferramenta que foi implementada para descompactar esse arquivo. Após a descompactação, foram processados somente os dados referentes a corrida e sono.

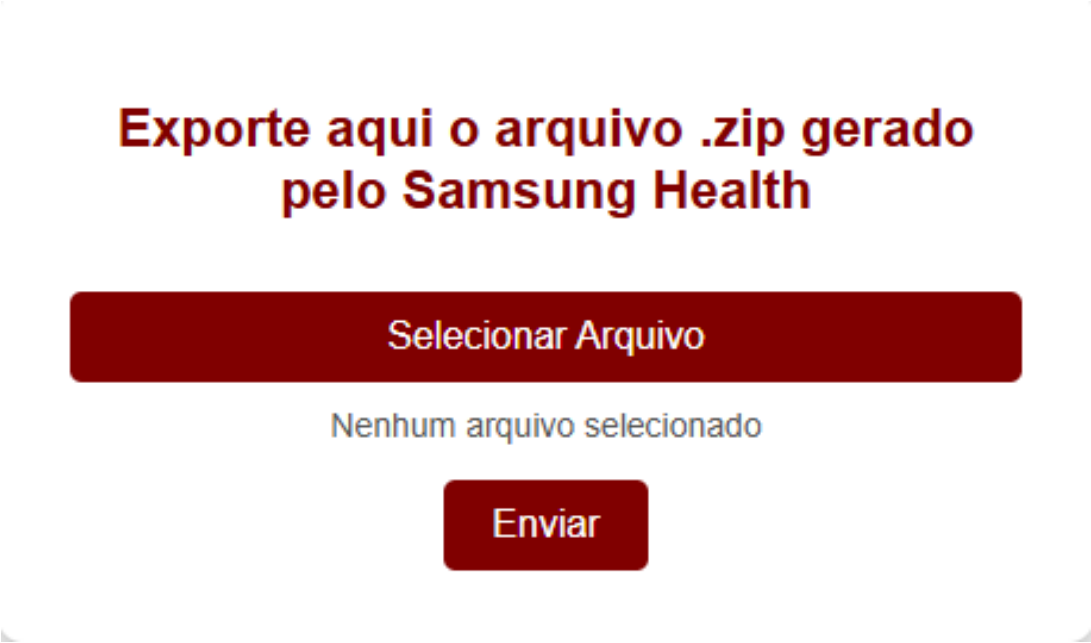

Figura 4.2: Interface para envio dos dados exportados

Na etapa de **processamento** os dados foram cuidadosamente examinados para identificar e corrigir possíveis erros, inconsistências ou lacunas. Isso incluiu a verificação da integridade dos dados coletados pelo *smartwatch*, a identificação de *outliers* e o uso de uma API para coletar a temperatura da data, local e hora que a atividade física foi realizada.

Assim, na etapa final, os **resultados** são apresentados ao usuário com a estimativa do tempo necessário para percorrer a distância inserida. Essa previsão é gerada pelo algoritmo de machine learning, que considera variáveis como temperatura no momento

da atividade, distância percorrida e outros dados coletados pelo *smartwatch*.

Em resumo, os dados coletados foram integrados e padronizados para garantir a uniformidade e consistência necessárias para a análise. Após a limpeza, foram aplicadas técnicas de pré-processamento, como normalização para preparar os dados para a construção dos modelos de Machine Learning.

4.1 Modelagem Preditiva para Corrida

O objetivo desta seção é apresentar a abordagem metodológica utilizada para a construção de modelos preditivos capazes de estimar o tempo necessário para percorrer uma determinada distância com base nos dados coletados. A modelagem preditiva busca identificar padrões e tendências nos registros de corrida, permitindo compreender os fatores que influenciam o desempenho dos corredores. A partir dessa análise, espera-se fornecer previsões mais precisas para que os atletas possam ajustar seu treinamento de forma estratégica a partir da estimativa de tempo retornada pelo modelo.

4.1.1 Algoritmos de Machine Learning Utilizados

Neste estudo, foram avaliados seis algoritmos de Machine Learning para a detecção de padrões nos dados de corrida: *Gradient Boosting, Support Vector Regression (SVR), K-Nearest Neighbors (KNN), Decision Tree, Random Forest e Linear Regression.* O objetivo foi prever o tempo estimado que um atleta amador levaria para percorrer uma determinada distância, informada pelo próprio usuário na interface do sistema.

Cada modelo foi treinado e testado para avaliar seu desempenho na tarefa de previsão. Durante a etapa de treinamento, foi necessária a padronização dos dados para evitar problemas em modelos sensíveis à escala. Além disso, foram realizados ajustes de hiper parâmetros para otimizar o desempenho dos algoritmos e garantir previsões mais precisas.

Os modelos utilizados possuem abordagens distintas para identificar padrões nos dados. O **Gradient Boosting** constrói modelos sequenciais de árvores de decisão, onde cada nova árvore tenta corrigir os erros das anteriores, proporcionando alta precisão, mas

exigindo maior capacidade computacional (GRADIENT..., 2024). O **Support Vector Regression (SVR)** busca um hiperplano ótimo para minimizar os erros dentro de uma margem definida, sendo útil para capturar relações não lineares nos dados (SCIKIT-LEARN, 2024d). Já o **K-Nearest Neighbors (KNN)** faz previsões com base na média dos valores de "K"vizinhos mais próximos, o que pode funcionar bem para padrões locais, mas é sensível a ruídos e dados fora do padrão (SCIKIT-LEARN, 2024b).

O **Decision Tree** estrutura suas decisões em uma hierarquia de regras, sendo um modelo interpretável, porém suscetível ao *overfitting* quando aplicado a dados complexos (SCIKIT-LEARN, 2024a). O **Random Forest** mitiga esse problema ao combinar múltiplas árvores de decisão treinadas em subconjuntos aleatórios dos dados, resultando em maior robustez e precisão (SCIKIT-LEARN, 2024c). Por fim, a **Linear Regression** ajusta uma relação linear entre as variáveis de entrada e o tempo de corrida, sendo eficaz para padrões mais simples e de fácil interpretação.

Outra estratégia adotada para aprimorar a precisão das previsões foi a correção de valores irreais. Essa abordagem visa ajustar o modelo sempre que o erro for 30% superior à média dos erros reais, permitindo a mitigação de *outliers* e a eliminação de previsões excessivamente distantes da realidade.

4.1.2 Avaliação e Métricas

Para avaliar a eficácia dos modelos de Machine Learning, foram utilizadas três métricas principais:

Coeficiente de Determinação (R^2): Mede o quanto da variação dos dados é explicada pelo modelo. Valores mais próximos de 1 indicam previsões mais precisas.

Erro Quadrático Médio da Raiz (RMSE): Representa a média das diferenças ao quadrado entre os valores previstos e os reais, penalizando erros maiores de forma mais significativa. Quanto menor o RMSE, melhor o desempenho do modelo.

Erro Médio Absoluto (MAE): Mede a diferença absoluta média entre os valores previstos e reais, sendo útil para interpretar os erros em unidades da variável alvo.

4.1.3 Resultados

Os modelos foram avaliados utilizando uma abordagem de divisão temporal dos dados, onde 80% das amostras mais antigas foram utilizadas para o treinamento e 20% mais recentes para o teste. Essa escolha se justifica pelo fato de que o desempenho de um corredor pode variar ao longo do tempo, e o objetivo do estudo é prever tempos futuros com base no histórico prévio. Dessa forma, evita-se o vazamento de dados e garante-se que o modelo reflita um cenário real de previsão.

Espera-se que os algoritmos de regressão sejam capazes de identificar tendências nos dados, permitindo a previsão do tempo necessário para percorrer uma determinada distância. Além disso, a análise dos erros nas previsões pode fornecer *insights* sobre a variabilidade no desempenho dos corredores e fatores que influenciam suas corridas.

Os resultados incluem a avaliação do desempenho dos modelos com base em métricas como R^2, *RMSE e MAE*, destacando sua precisão na previsão de tempos de corrida. Foi dada atenção especial à interpretação dos resultados e à aplicabilidade prática dos *insights* obtidos, visando oferecer um suporte inteligente para corredores não profissionais melhorarem seu desempenho e bem-estar.

4.2 Descrição dos dados

De modo geral, após o usuário fazer o *upload* do arquivo compactado que o *Samsung Health* gera, é feito todo o processamento dos dados. Na etapa de processamento são selecionados somente os arquivos csv que continham informações sobre a atividade física e sono.

Após toda essa etapa de manipulação dos dados obtivemos um conjunto de teste com 198 entradas, sendo 78 entradas numa faixa entre 3000 e 5000 metros. A seguir, para cada variável obtivemos a estatística descritiva conforme mostrada na Tabela 4.1.

Tabela 4.1: Estatísticas das variáveis

Variável	Média	Desvio Padrão	Mínimo	25%	Mediana	75%	Máximo
com.samsung.health.exercise.distance (m)	2313.21	1720.54	0	590.73	2291.15	3727.28	9162.56
com.samsung.health.exercise.mean_speed(mm/ms)	22914970	11454550	0	20265230	29002620	30911500	39162160
com.samsung.health.exercise.duration (min:s)	15:19	9:52	0:11	8:03	15:19	20:15	71:13
com.samsung.health.exercise.mean_heart_rate (bpm)	160	18	63	160	165	170	182
sleep_duration (min)	363	137	28	269	420	463	588
temperature_2m (°C)	17	4	6	15	18	20	27
relative_humidity_2m (%)	93	7	54	91	95	97	100
sleep_score	68	15	28	60	71	80	96
mental_recovery (%)	67	16	5	60	67	77	95

A Tabela 4.1 apresenta as estatísticas descritivas das variáveis numéricas utilizadas. Na etapa de processamento dos dados a *distância* veio em milímetros, o *tempo* em milissegundos, *velocidade média* em milímetros/milissegundos e o tempo de sono em minutos. Esse último, ao treinar o modelo,foi feita uma conversão em milissegundos para garantir que estejamos trabalhando com a mesma unidade de medida durante a etapa de treinamento. Para melhor visualização dos dados, foi feito um ajuste para as para unidades de medida mais adequadas com o objetivo de melhorar a visualização dos dados. A distância foi convertida para metros, enquanto a duração foi transformada para o formato minutos:segundos. Já a variável *sleep_score* seu valor varia numa escala de 0 a 100. Apesar das medidas apresentarem valores médios e percentis que auxiliam na descrição dos dados, é importante ressaltar que as distribuições estão enviesadas, uma vez que a maior parte das distâncias contidas nesses dados estão contidas numa faixa entre 3000 e 5000 metros. Essa característica ressalta a necessidade de maior cautela na interpretação dos resultados, considerando possíveis correções ou ajustes para minimizar os efeitos do viés nas análises subsequentes.

Ao analisar a tabela, identificamos nove variáveis distintas, que podem ser classificadas em dois grupos: variáveis *individuais* e variáveis *coletivas*.

As variáveis individuais são aquelas específicas de cada indivíduo, ou seja, dependem exclusivamente de suas características e condições pessoais. Nesse grupo, incluem-se a *velocidade média*, *batimento cardíaco*, *tempo de sono*, *recuperação mental* e *pontuação de sono*.

Já as variáveis coletivas são aquelas que independem do indivíduo, estando relaci-

onadas a fatores externos. Nesse grupo, encontram-se a *distância*, *temperatura* e *umidade*.

A matriz de correlação apresentada na Figura 4.3 demonstra as relações bivariadas entre as variáveis analisadas neste estudo. As cores representam a intensidade e a direção da correlação: tons de vermelho indicam correlações positivas, tons de azul indicam correlações negativas e a intensidade da cor reflete a força da correlação. Os valores numéricos dentro de cada célula representam o coeficiente de correlação de Pearson, variando de -1 a +1.

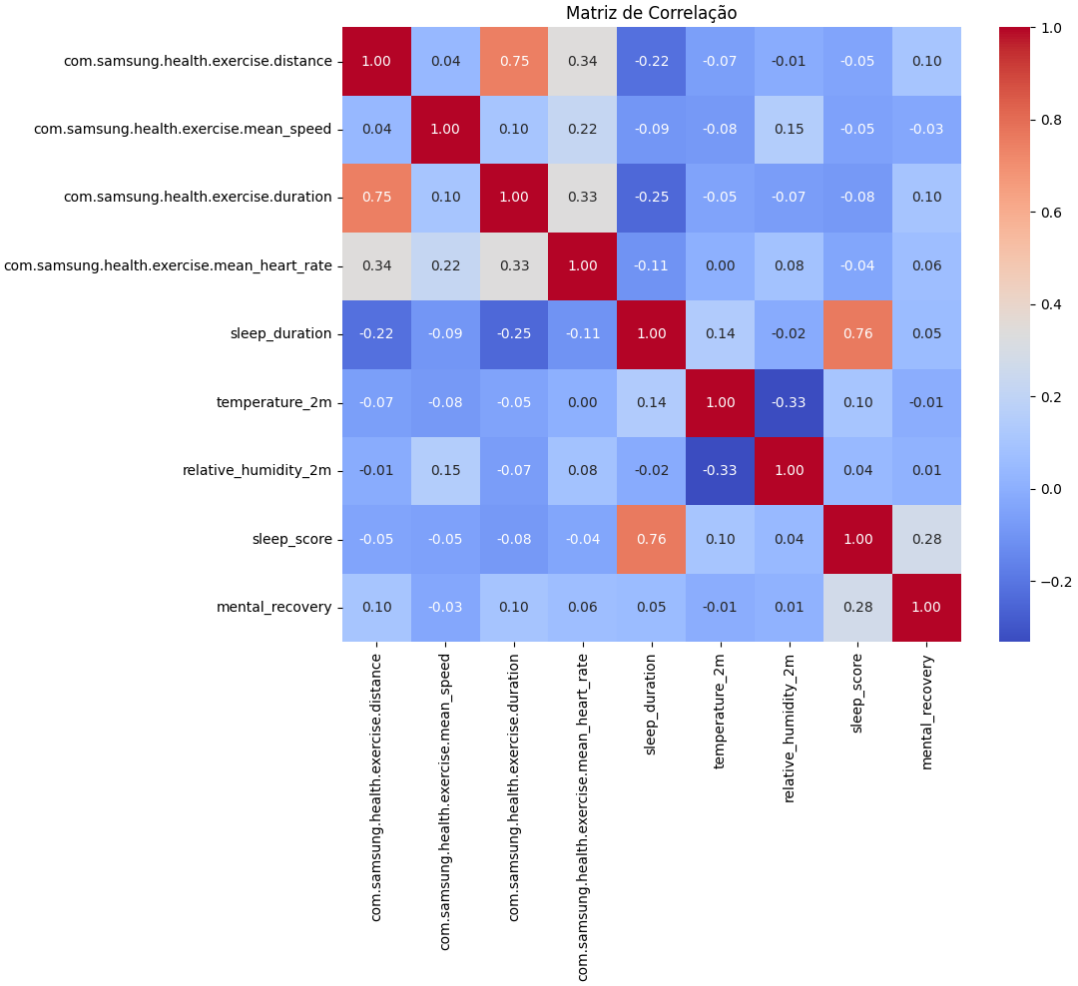

Figura 4.3: Interface para envio dos dados exportados

Ao analisar a matriz de correlação, destaca-se a presença de correlações negativas entre determinadas variáveis, como temperatura e duração do exercício. Conforme mencionado por (KNECHTLE BEAT; DI GANGI, 2019), é sabido que condições climáticas podem influenciar o desempenho de atletas, o que parece contradizer os resultados observados na matriz. No entanto, é importante ressaltar que a correlação indica apenas uma

relação estatística entre as variáveis, não implicando necessariamente em uma relação de causa e efeito. É plausível que outras variáveis não incluídas no estudo possam estar influenciando tanto o sono quanto o desempenho físico, o que poderia explicar as correlações negativas observadas. Portanto, a interpretação dos dados deve ser feita com cautela, considerando a complexidade das interações entre os fatores analisados.

4.3 Avaliação dos Modelos

Como dito anteriormente, os dados foram coletados de forma temporal uma vez que o desempenho de um corredor pode variar ao longo do tempo, e o objetivo do estudo é prever tempos futuros com base no histórico prévio. Diante disso, a Tabela 4.2 mostra os resultados obtidos após toda coleta dos dados, ajustes e treinamento dos modelos.

Tabela 4.2: Desempenho dos modelos de Machine Learning na previsão de tempo de corrida.

Nome do Modelo	**R^2**	**MAE (min:s)**	**RMSE (min:s)**
Gradient Boosting	0.58	2:21	3:58
SVR	-0.47	6:30	7:24
KNN	0.43	3:27	4:37
Decision Tree	0.60	2:12	3:52
Random Forest	0.58	2:20	3:57
Linear Regression	0.48	3:04	4:25

Ao analisar os resultados dos modelos de aprendizado de máquina aplicados, observamos que a *Árvore de Decisão (Decision Tree)* apresentou a melhor performance, tendo o maior Coeficiente de Determinação e menor MAE e RMSE, quando comparada aos outros modelos avaliados. A Árvore de Decisão tem a capacidade de identificar e modelar relações complexas e não lineares entre as variáveis.

Por outro lado, o *SVR* obteve o pior desempenho. Caso o conjunto de dados tenha grandes variações nas variáveis de entrada (como a distância), o SVR pode não ser capaz de capturar as flutuações de forma eficiente.

Considerando um cenário real de uso, as Tabelas 4.3, 4.4, 4.5 e 4.6 apresentam o

tempo previsto pelo modelo para diferentes distâncias inseridas pelo usuário como entrada no sistema de predição.

Tabela 4.3: Tempo estimado pelos modelos para uma distância de 5000 metros

Nome do Modelo	**Distância (m)**	**Tempo Estimado (min:s)**
Gradient Boosting	5000	37:51
SVR	5000	13:52
KNN	5000	18:10
Decision Tree	5000	28:32
Random Forest	5000	27:48
Linear Regression	5000	26:45

Tabela 4.4: Tempo estimado pelos modelos para uma distância de 6000 metros

Nome do Modelo	**Distância (m)**	**Tempo Estimado (min:s)**
Gradient Boosting	6000	37:51
SVR	6000	13:52
KNN	6000	18:10
Decision Tree	6000	28:32
Random Forest	6000	29:59
Linear Regression	6000	30:04

Tabela 4.5: Tempo estimado pelos modelos para uma distância de 7000 metros

Nome do Modelo	**Distância (m)**	**Tempo Estimado (min:s)**
Gradient Boosting	7000	37:51
SVR	7000	13:52
KNN	7000	21:44
Decision Tree	7000	28:32
Random Forest	7000	29:59
Linear Regression	7000	33:22

Tabela 4.6: Tempo estimado pelos modelos para uma distância de 10000 metros

Nome do Modelo	**Distância (m)**	**Tempo Estimado (min:s)**
Gradient Boosting	10000	37:51
SVR	10000	13:52
KNN	10000	28:30
Decision Tree	10000	28:32
Random Forest	10000	29:59
Linear Regression	10000	43:19

É importante destacar que, no conjunto de dados exportado pelo usuário, a maioria das corridas possui distâncias entre 3000 e 5000 metros. Isso significa que o modelo aprendeu sobre padrões específicos dentro desse intervalo, tornando suas previsões menos confiáveis para distâncias significativamente maiores ou menores.

Ao compararmos o tempo estimado pelos modelos com o tempo real de corrida do usuário para as distâncias analisadas e avaliarmos o erro obtido em cada algoritmo, observamos que o desempenho dos modelos foi mais preciso para distâncias próximas de 5.000 metros, faixa predominante no conjunto de dados.

Ao analisar o tempo estimado por cada algoritmo para diferentes distâncias inseridas, observamos que alguns modelos perdem a capacidade de generalização à medida que a distância aumenta. Esse comportamento indica que os modelos estão mais ajustados às faixas de distância predominantes no conjunto de dados, o que limita sua precisão para valores fora desse intervalo.

Para corridas de 5.000 metros, os modelos *Decision Tree, Random Forest* e *Linear Regression* apresentaram estimativas bastante próximas do tempo real do usuário. Em 6.000 metros, *Linear Regression* e *Random Forest* continuaram com boas previsões, enquanto o *Decision Tree*, embora com um erro ligeiramente maior, ainda manteve uma estimativa aceitável.

Em 7.000 metros, apenas o *Random Forest* conseguiu manter uma previsão precisa, sugerindo que esse modelo tem melhor capacidade de generalização para distâncias superiores às mais frequentes no conjunto de treino. Entretanto, ao analisarmos a previsão

para 10.000 metros, nenhum dos modelos conseguiu se adaptar bem ao tempo real, indicando que a capacidade preditiva dos algoritmos se torna menos confiável para distâncias significativamente superiores ao intervalo predominante nos dados de treino.

Essa limitação na capacidade preditiva dos modelos para distâncias muito maiores ou menores do que aquelas presentes na base de dados pode estar relacionada a diversos fatores. Um dos principais é o viés induzido pela distribuição dos dados de treino. Como a maioria das corridas registradas está entre 3.500 e 5.000 metros, os modelos ajustam seus parâmetros para prever com maior precisão dentro desse intervalo. No entanto, quando confrontados com distâncias significativamente diferentes, podem extrapolar padrões de forma inadequada, resultando em previsões menos confiáveis.

Outro fator relevante é a relação não necessariamente linear entre distância e tempo de corrida. Embora a tendência geral seja que tempos maiores estejam associados a distâncias maiores, o comportamento real pode ser influenciado por diversas variáveis, como fadiga, variação no ritmo e condições externas.

Além disso, modelos baseados em árvores, como *Random Forest* e *Decision Tree*, podem apresentar dificuldades para generalizar além do espaço amostral porque tomam decisões baseadas em divisões nos dados de treino. Se não houver amostras suficientes de corridas longas, essas decisões tornam-se menos precisas para distâncias superiores.

5 Considerações Finais e Trabalhos Futuros

Este trabalho explorou a aplicação de algoritmos de Machine Learning para prever o tempo de corrida de corredores não profissionais, utilizando dados extraídos do *Samsung Health*. Ao longo da pesquisa, foram analisados diferentes modelos de regressão, avaliando sua capacidade de generalização e precisão em diferentes distâncias. Os resultados obtidos mostraram que o *Random Forest* apresentou o melhor desempenho geral para distâncias próximas à faixa predominante no conjunto de dados. Além disso, identificamos limitações na capacidade dos modelos de extrapolar previsões para distâncias significativamente maiores, o que reforça a importância de possuir bastante variação dos dados do conjunto de treino.

Embora o modelo não tenha se adaptado tão bem em distâncias mais longas, este estudo demonstra seu potencial para causar um impacto positivo em atletas não profissionais, oferecendo uma ferramenta inteligente para estimar tempos de corrida com base em dados históricos. A análise dos modelos de Machine Learning permitiu identificar padrões relevantes de desempenho, auxiliando corredores na definição de metas mais realistas e estratégias de treino mais eficazes.

Ao ter uma estimativa precisa do tempo necessário para percorrer uma determinada distância, o atleta pode planejar sua corrida de forma estratégica, otimizando seu desempenho e evitando desgaste desnecessário. Por exemplo, um profissional de educação física, conhecendo previamente o tempo previsto para seu aluno completar a prova, pode estruturar um plano detalhado, indicando as zonas de frequência cardíaca ideais para cada trecho do percurso. Isso é especialmente útil para iniciantes, que frequentemente começam em um ritmo muito forte e acabam perdendo rendimento ao longo da corrida.

Com uma previsão bem ajustada, o atleta pode distribuir melhor sua energia, ajustando a intensidade do esforço em cada fase da prova. Isso possibilita manter um ritmo sustentável, evitar quedas bruscas de desempenho e potencializar a eficiência da corrida, resultando em um tempo final mais consistente e melhor aproveitamento fisiológico.

Além disso, ao comparar o tempo real com o tempo estimado, é possível levan-

tar hipóteses sobre fatores que possam ter influenciado o desempenho do atleta. Caso o tempo real tenha sido maior do que o previsto, isso pode indicar que o atleta enfrentou dificuldades involuntárias, como um suporte nutricional inadequado, desidratação ou baixos níveis de glicogênio muscular, afetando sua resistência e capacidade de manter o ritmo.

Outro fator relevante pode ser a fadiga acumulada, seja por treinos intensos anteriores, recuperação insuficiente ou até mesmo por questões externas, como estresse e qualidade do sono. Identificar essas variações permite ajustes estratégicos no planejamento de treinos, na alimentação e na recuperação, contribuindo para um desempenho mais consistente e eficiente ao longo do tempo.

Desse modo, a pesquisa reforça a importância da personalização dos modelos para diferentes perfis de corredores, destacando a necessidade de bases de dados mais diversificadas para melhorar a precisão das previsões. Os *insights* obtidos podem servir de base para o desenvolvimento de novas aplicações voltadas ao acompanhamento e otimização do desempenho esportivo, tornando a tecnologia um aliado valioso para atletas amadores em sua evolução no esporte.

Com base nesse trabalho e tudo que foi discutido podemos pensar em alguns pontos que podem ser usados em trabalhos futuros, visando melhorar ainda mais a precisão desse modelo:

- **Diversificar e aumentar a quantidade de dados:** Ao diversificar e aumentar a quantidade dos dados por meio da coleta de informações de diferentes corredores, é possível criar perfis de usuário que representem distintos níveis de condicionamento físico. Isso não apenas melhora a capacidade de generalização do modelo, tornando-o mais preciso para uma gama maior de atletas.

- **Uso de modelos mais avançados:** Testar a possibilidade de utilizar Aprendizagem Profunda (Deep Learning) para aumentar a precisão das previsões.

- **Integração com dispositivos vestíveis:** Estudar a possibilidade de uso de dados em tempo real de *smartwatches* para fornecer previsões mais dinâmicas e adaptáveis ao contexto da corrida.

- **Transferência de Aprendizado:** Como a coleta de dados de corredores amadores neste trabalho foi restrita, a Transferência de Aprendizado permite utilizar conhecimento extraído de bases maiores para melhorar a precisão das previsões, mesmo com poucos dados específicos do usuário. Além disso, essa abordagem pode reduzir o risco de *overfitting*, garantindo que o modelo consiga generalizar melhor.

Bibliografia

ALSAREII, S. A. et al. Physical activity monitoring and classification using machine learning techniques. In: *National Library of Medicine*. [S.l.]: National Library of Medicine, 2022.

BONAT, L. R. *Corridas de rua se tornam cada vez mais populares*. Comunicare, 2024. Disponível em: ⟨https://www.portalcomunicare.com.br/corridas-de-rua-se-tornam-cada-vez-mais-populares-confira-proximas-datas-em-curitiba/#:~:text=Segundo%20dados%20levantados%20pela%20Associa%C3%A7%C3%A3o,13%20milh%C3%B5es%20de%20corredores%20atualmente.⟩

BRUCE, C. *benefícios da corrida para a saúde*. Grupo Rede Dor, 2023. Disponível em: ⟨https://www.tuasaude.com/beneficios-da-corrida/⟩.

CIABATTONI, L. et al. Real-time mental stress detection based on smartwatch. In: *IEEE*. [S.l.]: 2017 IEEE International Conference on Consumer Electronics (ICCE), 2017. p. 110–111.

EYSENBACH, G. What is e-health? In: *Journal of medical Internet research*. Toronto, Canada: JMIR Publications, 2001. v. 3, n. 2, p. e833.

GRADIENT Boosting. Scikit-learn, 2024. Disponível em: ⟨https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting⟩.

International Organization for Standardization. *ISO/IEC 25010:2011 - Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*. International Organization for Standardization, 2011. Disponível em: ⟨https://iso25000.com/index.php/en/iso-25000-standards/iso-25010⟩.

KNECHTLE BEAT; DI GANGI, S. R. C. A. V. E. R. T. N. P. T. The role of weather conditions on running performance in the boston marathon from 1972 to 2018. In: . Zurich, Switzerland: Public Library of Science (PLoS), 2019.

LOTTENBERG, C. *Pandemia fez brasileiros apostarem em hábitos mais saudáveis*. Veja, 2021. Disponível em: ⟨https://veja.abril.com.br/coluna/coluna-claudio-lottenberg/pandemia-fez-brasileiros-apostarem-em-habitos-mais-saudaveis⟩.

MANTZIOS, K. et al. Effects of weather parameters on endurance running performance: Discipline-specific analysis of 1258 races. In: *National Library of Medicine*. [S.l.]: Pub Med, 2021.

PIEDRAS, J. *Nesta segunda é comemorado o Dia do Esporte Amador*. Orla Rio, 2021. Disponível em: ⟨https://orlario.com.vc/home/nesta-segunda-e-comemorado-o-dia-do-esporte-amador/⟩.

SCIKIT-LEARN. *Decision Trees*. Scikit-learn, 2024. Accessed: 2024-03-10. Disponível em: ⟨https://scikit-learn.org/stable/modules/tree.html⟩.

SCIKIT-LEARN. *K-Nearest Neighbors (KNN)*. Scikit-learn, 2024. Accessed: 2024-03-10. Disponível em: ⟨https://scikit-learn.org/stable/modules/neighbors.html#k-neighbors-classification⟩.

SCIKIT-LEARN. *Random Forest*. Scikit-learn, 2024. Accessed: 2024-03-10. Disponível em: ⟨https://scikit-learn.org/stable/modules/ensemble.html#random-forest⟩.

SCIKIT-LEARN. *Support Vector Regression (SVR)*. Scikit-learn, 2024. Accessed: 2024-03-10. Disponível em: ⟨https://scikit-learn.org/stable/modules/svm.html#regression⟩.

TENORIO, G. *Autocuidado em tempos de pandemia*. Veja, 2020. Disponível em: ⟨https://saude.abril.com.br/especiais/autocuidado-em-tempos-de-pandemia⟩.

VERDE, C. *40Disponível em: ⟨https://cidadeverde.com/noticias/397042/40-dos-brasileiros-estao-mais-preocupados-com-a-saude-fisica-diz-pesquisa⟩.*

WEISER, M. The computer for the 21st century. In: . [S.l.]: Acm Digital Library, 1991.

World Health Organization. *Fifty-eighth World Health Assembly: WHA58.28 - Public Health, Innovation, and Intellectual Property*. World Health Organization, 2005. Disponível em: ⟨https://apps.who.int/gb/ebwha/pdf_files/WHA58/WHA58_28-en.pdf⟩.

YEUNG, T. *O Que É Computação na Borda?* Nvidia blog, 2023. Disponível em: ⟨https://blog.nvidia.com.br/blog/o-que-e-computacao-na-borda/⟩.