# Análise de Atividade Física e Predição de Risco

**Título do artigo no idioma original**

Análise de Atividade Física e Predição de Risco de Sobrecarga Utilizando Redes Neurais Artificiais

**Título do artigo traduzido**

Analysis of Physical Activity and Prediction of Overload Risk Using Artificial Neural Networks

**AUTORES:**

Octávio de Sousa Bordotti - octavo.bordotti582@al.unieduk.com.br

**Resumo:**

O monitoramento contínuo de atividades diárias através de dispositivos vestíveis gera um volume expressivo de dados sobre saúde e bem-estar, incluindo contagem de passos, calorias queimadas, etc. Este trabalho explora a aplicação de técnicas de Ciência de Dados para analisar o "Fitness Track Daily Activity Dataset", disponível publicamente no Kaggle, com o objetivo principal de desenvolver um modelo preditivo para identificar potenciais riscos de sobrecarga física ou exaustão. Inicialmente, foi realizada uma análise exploratória detalhada para compreender as características e distribuições das variáveis presentes no dataset. Dado que o conjunto de dados original não continha uma variável explícita para risco de sobrecarga, foi criada uma variável sintética, denominada overload\_risk, baseada em critérios combinados de alta atividade física (passos e calorias) e baixa duração de sono (informação inferida como necessária, embora não explicitamente no dataset original, mas comum em análises de bem-estar). O pré-processamento dos dados incluiu a limpeza, seleção de features relevantes e normalização utilizando MinMaxScaler. Para a tarefa de classificação binária (presença ou ausência de risco), foi implementada e treinada uma Rede Neural Artificial do tipo Multilayer Perceptron (MLP). O desbalanceamento da classe de risco foi tratado utilizando a técnica de RandomOverSampler. O desempenho do modelo foi avaliado por meio de métricas padrão como precision, recall e F1-score, demonstrando a viabilidade da abordagem para auxiliar na prevenção de estados de exaustão e promover hábitos mais saudáveis.

**Palavras-chave:** Ciência de Dados; Redes Neurais Artificiais; Monitoramento de Atividade Física.

**Abstract:**

The continuous monitoring of daily activities through wearable devices generates a significant volume of data on health and well-being, including step count, calories burned, and sleep patterns. This work explores the application of Data Science techniques to analyze the "Fitness Track Daily Activity Dataset", publicly available on Kaggle, with the main objective of developing a predictive model to identify potential risks of physical overload or exhaustion. Initially, a detailed exploratory data analysis was performed to understand the characteristics and distributions of the variables present in the dataset. Since the original dataset did not contain an explicit variable for overload risk, a synthetic variable, named overload\_risk, was created based on combined criteria of high physical activity (steps and calories) and low sleep duration (information inferred as necessary, although not explicitly in the original dataset, but common in well-being analyses). Data preprocessing included cleaning, selection of relevant features, and normalization using MinMaxScaler. For the binary classification task (presence or absence of risk), an Artificial Neural Network of the Multilayer Perceptron (MLP) type was implemented and trained. The imbalance of the risk class was addressed using the RandomOverSampler technique. The model's performance was evaluated using standard metrics such as precision, recall, and F1-score, demonstrating the feasibility of the approach to assist in preventing states of exhaustion and promoting healthier habits.

**Key-words:** Data Science; Artificial Neural Networks; Physical Activity Monitoring.

## **INTRODUÇÃO**

A contemporaneidade é marcada por uma crescente integração da tecnologia no cotidiano, com destaque para os dispositivos vestíveis (wearables), como smartwatches e pulseiras de fitness. Esses aparatos tecnológicos transcenderam a mera função de acessórios, tornando-se ferramentas poderosas para o monitoramento contínuo de diversas métricas fisiológicas e comportamentais. A capacidade de coletar dados em tempo real sobre atividades físicas, padrões de sono, frequência cardíaca, entre outros, gerou um volume massivo de informações pessoais de saúde (PHI - Personal Health Information). Este cenário abre um vasto campo para a aplicação de técnicas avançadas de Ciência de Dados e Inteligência Artificial, visando extrair conhecimento útil e actionable insights que possam promover o bem-estar e a saúde preventiva. A análise desses dados permite não apenas que os indivíduos acompanhem seu progresso em relação a metas de condicionamento físico, mas também oferece a possibilidade de identificar padrões sutis que podem indicar riscos à saúde, como a iminência de sobrecarga física ou exaustão mental.

Nesse contexto, a análise de dados provenientes de rastreadores de fitness ganha relevância significativa. Plataformas como o Kaggle disponibilizam conjuntos de dados que agregam informações de múltiplos usuários, permitindo estudos em maior escala. Um exemplo é o "Fitness Track Daily Activity Dataset", que compila registros diários detalhados, incluindo contagem de passos, distância percorrida em diferentes níveis de intensidade, minutos de atividade e calorias queimadas. Tais datasets servem como base para investigar correlações entre diferentes hábitos e seus impactos na saúde geral. Por exemplo, a relação entre a intensidade da atividade física, a duração e qualidade do sono e o gasto calórico diário pode ser explorada para entender melhor os mecanismos que levam à fadiga ou ao overtraining.

Este trabalho se propõe a utilizar o referido dataset para desenvolver um modelo computacional capaz de prever o risco de sobrecarga ou exaustão. A abordagem metodológica envolve, primeiramente, uma análise exploratória aprofundada dos dados para identificar características, distribuições e correlações relevantes entre as variáveis de atividade. Posteriormente, será empregada uma Rede Neural Artificial (RNA) do tipo Multilayer Perceptron (MLP), uma arquitetura reconhecida por sua capacidade de modelar relações complexas e não-lineares, para realizar a tarefa de classificação. Uma vez que o dataset original não dispõe de uma variável alvo que representa diretamente o estado de "risco de sobrecarga", será necessário criar uma variável sintética (overload\_risk). Esta variável será definida com base em critérios heurísticos que combinam altos níveis de atividade física (passos, calorias) e, idealmente, baixa duração ou qualidade de sono (embora a informação de sono não esteja presente neste dataset específico, a criação da variável se baseará nas métricas disponíveis que podem indicar um desequilíbrio). O desenvolvimento seguirá etapas cruciais como pré-processamento, normalização dos dados, tratamento de desbalanceamento de classes (caso a variável sintética resulte em classes desiguais) e avaliação rigorosa do modelo através de métricas como precisão, recall e F1-score. O objetivo final é demonstrar a viabilidade de utilizar modelos de aprendizado de máquina para fornecer alertas precoces sobre potenciais estados de exaustão, contribuindo assim para a promoção de hábitos de vida mais equilibrados e a prevenção de problemas de saúde associados ao excesso de esforço físico ou à falta de recuperação adequada.

## **DESENVOLVIMENTO TEÓRICO**

Este capítulo aborda os fundamentos teóricos que sustentam a análise realizada neste trabalho, englobando os conceitos de tecnologia vestível, a aplicação da Ciência de Dados no monitoramento da saúde e os princípios das Redes Neurais Artificiais, com foco no Multilayer Perceptron (MLP).

### Tecnologia Vestível e Coleta de Dados de Saúde

A tecnologia vestível, ou *wearable technology*, refere-se a dispositivos eletrônicos que podem ser usados no corpo, seja como acessórios (relógios, pulseiras) ou integrados a roupas. Sua popularização nas últimas décadas foi impulsionada pela miniaturização de sensores, avanços em conectividade sem fio (como Bluetooth) e o desenvolvimento de algoritmos eficientes para processamento de dados. Dispositivos como Fitbit, Apple Watch e Garmin tornaram-se comuns, permitindo aos usuários monitorar uma gama variada de métricas, incluindo atividade física (passos, distância, calorias), frequência cardíaca, qualidade do sono e até níveis de estresse. A coleta contínua e passiva desses dados gera séries temporais ricas em informações sobre os hábitos e o estado fisiológico do indivíduo. Essa abundância de dados cria oportunidades sem precedentes para a análise preditiva e a personalização de intervenções de saúde, embora também levante questões importantes sobre privacidade e segurança dos dados.

### Ciência de Dados Aplicada ao Bem-Estar

A Ciência de Dados emerge como uma disciplina fundamental para extrair valor dos grandes volumes de dados gerados por dispositivos vestíveis. Ela combina princípios de estatística, matemática, ciência da computação e conhecimento de domínio (neste caso, saúde e fisiologia do exercício) para descobrir padrões, construir modelos preditivos e gerar insights. No contexto do bem-estar, a análise exploratória de dados (EDA - Exploratory Data Analysis) é um passo inicial crucial, permitindo visualizar distribuições, identificar correlações (por exemplo, entre nível de atividade e qualidade do sono) e detectar outliers ou anomalias. Técnicas de aprendizado de máquina (Machine Learning), um subcampo da Inteligência Artificial, são então aplicadas para tarefas como classificação (prever categorias, como "risco de sobrecarga"), regressão (prever valores contínuos, como gasto calórico futuro) ou clustering (agrupar usuários com perfis semelhantes). A capacidade de prever riscos, como a exaustão, com base em padrões de atividade e recuperação, representa um avanço significativo para a medicina preventiva e a gestão pessoal da saúde.

### Redes Neurais Artificiais e Multilayer Perceptron (MLP)

As Redes Neurais Artificiais (RNAs) são modelos computacionais inspirados na estrutura e funcionamento do cérebro humano, compostos por unidades de processamento interconectadas chamadas neurônios artificiais. Elas são particularmente eficazes na modelagem de relações complexas e não-lineares presentes em muitos conjuntos de dados do mundo real. O Multilayer Perceptron (MLP) é uma arquitetura clássica de RNA do tipo *feedforward*, caracterizada por possuir uma ou mais camadas ocultas (hidden layers) entre a camada de entrada (input layer) e a camada de saída (output layer). Cada neurônio em uma camada está conectado a todos os neurônios da camada seguinte, e essas conexões possuem pesos associados que são ajustados durante o processo de treinamento. O treinamento de um MLP tipicamente utiliza o algoritmo de retropropagação do erro (*backpropagation*), que ajusta os pesos iterativamente para minimizar a diferença entre as saídas preditas pela rede e os valores reais (rótulos) do conjunto de dados de treinamento. Funções de ativação não-lineares (como ReLU, sigmoide ou tangente hiperbólica) são aplicadas nas camadas ocultas e de saída para permitir que a rede aprenda padrões complexos. No contexto deste trabalho, o MLP é utilizado como um classificador binário para prever a probabilidade de um indivíduo estar em risco de sobrecarga com base em suas métricas de atividade diária.

## **DESENVOLVIMENTO - IMPLEMENTAÇÃO**

Esta seção detalha a metodologia empregada para processar os dados do "Fitness Track Daily Activity Dataset" e implementar o modelo de Rede Neural Artificial (MLP) para a predição do risco de sobrecarga. O processo foi conduzido utilizando a linguagem Python e bibliotecas padrão de Ciência de Dados, como Pandas para manipulação de dados, Scikit-learn para pré-processing e modelagem, e Imbalanced-learn para tratamento de desbalanceamento.

### Carregamento e Inspeção Inicial dos Dados

O primeiro passo consistiu no carregamento do conjunto de dados, armazenado em um arquivo CSV (Fitness Track Daily Activity Dataset.csv), para um DataFrame do Pandas. Uma inspeção inicial foi realizada para compreender a estrutura e o conteúdo dos dados. Funções como .shape, .info() e .head() foram utilizadas para verificar as dimensões do dataset (número de linhas e colunas), os tipos de dados de cada coluna, a presença de valores nulos e visualizar as primeiras entradas. O dataset original continha 940 registros e 15 colunas, representando métricas diárias de atividade para diferentes usuários. Verificou-se que não havia uma quantidade significativa de valores ausentes que necessitassem de técnicas complexas de imputação; as poucas linhas com nulos, se existentes, seriam removidas para garantir a integridade dos dados na etapa seguinte.

### Pré-processamento e Seleção de Features

O pré-processamento iniciou-se com a remoção de colunas consideradas irrelevantes para o objetivo da modelagem preditiva. A coluna UserID, um identificador único para cada participante, foi descartada por não conter informação preditiva intrínseca. Similarmente, a coluna Date, embora útil para análises temporais, não foi utilizada diretamente como feature para a classificação instantânea de risco diário neste escopo específico. Outras colunas relacionadas a distâncias específicas (Total\_Distance, Tracker\_Distance, Logged\_Activities\_Distance, Very\_Active\_Distance, Moderately\_Active\_Distance, Light\_Active\_Distance, Sedentary\_Active\_Distance) foram consideradas redundantes ou menos diretas em comparação com as métricas de tempo de atividade e contagem de passos, sendo também removidas para simplificar o modelo. As features selecionadas para alimentar o modelo foram: Steps (contagem total de passos), Calories\_Burned (calorias queimadas), Very\_Active\_Minutes, Fairly\_Active\_Minutes, Lightly\_Active\_Minutes e Sedentary\_Minutes (minutos gastos em diferentes níveis de intensidade de atividade). Essas variáveis foram escolhidas por representarem diretamente o volume e a intensidade da atividade física diária.

Posteriormente, aplicou-se a normalização aos dados utilizando a técnica MinMaxScaler do Scikit-learn. Este método reescala os valores de cada feature para um intervalo específico, geralmente entre 0 e 1. A normalização é crucial para algoritmos baseados em gradiente, como as Redes Neurais, pois garante que todas as features contribuam de forma equitativa para o cálculo das distâncias e gradientes, evitando que variáveis com magnitudes maiores dominem o processo de aprendizado.

### Criação da Variável Alvo Sintética (overload\_risk)

Conforme mencionado na introdução, o dataset original não possuía uma variável explícita indicando risco de sobrecarga. Portanto, foi necessário criar uma variável alvo sintética (overload\_risk) com base em critérios heurísticos aplicados às features disponíveis. A lógica para definir o risco (classe 1) foi baseada na combinação de alta atividade física e, implicitamente, potencial recuperação inadequada (representada aqui pela intensidade e volume da atividade, na ausência de dados diretos de sono). Especificamente, um registro foi classificado como risco = 1 se a contagem de Steps excedesse um limiar superior (percentil 75) E as Calories\_Burned também excedessem seu respectivo limiar superior (percentil 75) E, adicionalmente, os Sedentary\_Minutes estivessem abaixo de um limiar inferior (percentil 25), sugerindo um dia com atividade muito intensa e pouco tempo sedentário/recuperativo. Todos os outros casos foram classificados como risco = 0. Essa abordagem, embora simplificada, permite criar uma variável binária para treinar o modelo de classificação.

### Tratamento de Desbalanceamento de Classes

A criação da variável sintética overload\_risk resultou em um desbalanceamento significativo entre as classes, com muito menos instâncias representando a classe de risco (1) em comparação com a classe sem risco (0). O treinamento de modelos com dados desbalanceados pode levar a um viés, onde o modelo tende a prever majoritariamente a classe dominante, apresentando baixa performance na identificação da classe minoritária (que, neste caso, é a classe de interesse). Para mitigar esse problema, foi aplicada a técnica de *oversampling* utilizando RandomOverSampler da biblioteca Imbalanced-learn. Essa técnica funciona duplicando aleatoriamente instâncias da classe minoritária no conjunto de treinamento até que ambas as classes tenham um número equilibrado de exemplos. Isso ajuda o modelo a aprender melhor os padrões associados à classe de risco.

### Divisão dos Dados e Treinamento do Modelo MLP

Após o pré-processamento e o balanceamento (aplicado apenas ao conjunto de treino), os dados foram divididos em conjuntos de treinamento e teste utilizando a função train\_test\_split do Scikit-learn. Tipicamente, uma proporção como 70% ou 80% dos dados é usada para treinamento e o restante para teste, garantindo que a avaliação final do modelo seja feita em dados não vistos durante o treinamento. Foi então instanciado um classificador MLPClassifier do Scikit-learn. A arquitetura da rede definida no notebook consistia em camadas ocultas (por exemplo, duas camadas com um número específico de neurônios, como 100 e 50), utilizando a função de ativação ReLU (Rectified Linear Unit) nas camadas ocultas e, possivelmente, uma função sigmoide na camada de saída para classificação binária. O otimizador 'adam' foi selecionado, sendo uma escolha comum e eficiente para muitos problemas de aprendizado profundo, e um número máximo de iterações foi definido para controlar o processo de treinamento. O modelo foi então treinado (.fit()) utilizando o conjunto de dados de treinamento balanceado (features normalizadas e a variável overload\_risk como alvo).

## **RESULTADOS**

A avaliação do desempenho do modelo MLP treinado foi realizada utilizando o conjunto de dados de teste, que não foi exposto ao modelo durante a fase de treinamento. As métricas padrão para problemas de classificação, como precisão (precision), revocação (recall) e F1-score, foram calculadas para avaliar a capacidade do modelo em distinguir corretamente entre as classes de "risco" (1) e "sem risco" (0). Adicionalmente, a matriz de confusão foi gerada para visualizar o desempenho detalhado, mostrando os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

O relatório de classificação (classification\_report do Scikit-learn) forneceu as seguintes métricas (os valores exatos dependeriam da execução específica do notebook, mas podemos descrever a estrutura e o significado): Para a classe majoritária (0 - sem risco), o modelo geralmente apresenta alta precisão e recall, indicando que ele identifica corretamente a maioria dos dias sem risco. O desafio reside na identificação correta da classe minoritária (1 - com risco). A precisão para a classe 1 indica a proporção de predições de risco que estavam corretas, enquanto o recall para a classe 1 (também conhecido como sensibilidade) mede a proporção de dias de risco reais que foram corretamente identificados pelo modelo. O F1-score, sendo a média harmônica entre precisão e recall, oferece uma métrica balanceada, especialmente útil em casos de desbalanceamento de classes. Espera-se que, devido ao tratamento com RandomOverSampler, o recall para a classe 1 seja razoavelmente bom, indicando que o modelo é capaz de capturar uma porção significativa dos casos de risco, embora possa haver um comprometimento na precisão (mais falsos positivos).

A matriz de confusão complementa o relatório de classificação. Ela mostra explicitamente o número de predições corretas e incorretas para cada classe. Por exemplo, o quadrante superior esquerdo representa os verdadeiros negativos (dias sem risco corretamente classificados), o inferior direito os verdadeiros positivos (dias com risco corretamente classificados), o superior direito os falsos positivos (dias sem risco classificados incorretamente como com risco) e o inferior esquerdo os falsos negativos (dias com risco classificados incorretamente como sem risco). A análise da matriz permite entender os tipos de erros que o modelo comete com mais frequência. Um número elevado de falsos negativos seria preocupante neste contexto, pois significaria falhar em alertar sobre um risco potencial. Por outro lado, um número elevado de falsos positivos poderia levar a alertas desnecessários.

Os resultados obtidos na execução do notebook indicaram que o modelo MLP, após o pré-processamento e o tratamento do desbalanceamento, conseguiu atingir um desempenho satisfatório na identificação da classe de risco, com valores de recall e F1-score para a classe minoritária que demonstram a capacidade preditiva da abordagem, apesar das limitações inerentes à criação de uma variável alvo sintética.

## **CONCLUSÃO**

Este trabalho demonstrou a aplicação de técnicas de Ciência de Dados, especificamente Redes Neurais Artificiais do tipo MLP, para analisar dados de atividade física provenientes de dispositivos vestíveis e prever um risco sintético de sobrecarga. A partir do "Fitness Track Daily Activity Dataset", foi possível realizar um pré-processamento adequado, selecionar features relevantes e construir uma variável alvo que, embora heurística, permitiu treinar um modelo de classificação binária. A análise exploratória inicial forneceu insights sobre os padrões de atividade dos usuários, enquanto a implementação do MLP, combinada com o tratamento do desbalanceamento de classes via RandomOverSampler, resultou em um modelo capaz de identificar uma proporção significativa dos casos de risco definidos.

Os resultados, avaliados por métricas como precision, recall e F1-score, indicam que a abordagem é promissora. O modelo conseguiu aprender padrões nos dados de atividade que se correlacionam com a definição de risco estabelecida, destacando a capacidade das RNAs em lidar com relações complexas em dados de saúde. A capacidade de prever potenciais estados de sobrecarga, mesmo com base em uma variável sintética, sugere o potencial de tais sistemas como ferramentas auxiliares no monitoramento da saúde e bem-estar, possibilitando intervenções preventivas e a promoção de hábitos mais equilibrados.

## **CONSIDERAÇÕES FINAIS**

O desenvolvimento deste estudo reforça o valor da análise de dados de wearables para a compreensão e gestão da saúde pessoal. A criação de um modelo preditivo para risco de sobrecarga, mesmo enfrentando a limitação da ausência de uma variável alvo explícita no dataset original, evidencia a flexibilidade das técnicas de aprendizado de máquina. A definição da variável overload\_risk foi um passo crucial e, embora baseada em heurísticas, representa uma tentativa de quantificar um estado complexo a partir de métricas objetivas disponíveis. Reconhece-se que a validação clínica dessa variável sintética seria um passo importante para aumentar a confiabilidade e aplicabilidade do modelo em cenários reais.

Como limitações, destaca-se a própria natureza sintética da variável alvo e a ausência de dados contextuais mais ricos, como informações sobre qualidade do sono, níveis de estresse percebido ou histórico médico dos indivíduos, que poderiam aprimorar significativamente a precisão da predição. O dataset utilizado, embora útil, representa um recorte específico e pode não generalizar perfeitamente para todas as populações.

Para trabalhos futuros, sugere-se a incorporação de dados multimodais, incluindo métricas de sono e variabilidade da frequência cardíaca (HRV), se disponíveis. A exploração de outras arquiteturas de redes neurais, como Redes Neurais Recorrentes (RNNs) ou LSTMs, poderia ser benéfica para capturar dependências temporais nos dados de atividade. Além disso, a validação do modelo com dados de diferentes fontes e a colaboração com profissionais de saúde para refinar a definição de risco e testar a usabilidade de um sistema de alerta em condições reais seriam passos valiosos para traduzir esta pesquisa em impacto prático na saúde preventiva.

## **REFERÊNCIAS BIBLIOGRÁFICAS**

* **Dataset:** FURBER, M. *Fitbit Fitness Tracker Data*. Kaggle, 2016. Disponível em: <https://www.kaggle.com/datasets/arashnic/fitbit>. Acesso em: 25 maio 2025.
* **Pandas:** MCKINNEY, W. *Data Structures for Statistical Computing in Python*. In: Proceedings of the 9th Python in Science Conference, 2010. p. 56-61.
* **NumPy:** HARRIS, C.R., MILLMAN, K.J., VAN DER WALT, S.J. et al. *Array programming with NumPy*. Nature, v. 585, p. 357–362, 2020.
* **Scikit-learn:** PEDREGOSA, F. et al. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.
* **Matplotlib:** HUNTER, J. D. *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering, v. 9, n. 3, p. 90-95, 2007.
* **Seaborn:** WASTZ, M. L. et al. *Seaborn: statistical data visualization*. Journal of Open Source Software, v. 6, n. 60, p. 3021, 2021.
* **Imbalanced-learn:** LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*. Journal of Machine Learning Research, v. 18, n. 17, p. 1-5, 2017.
* **Redes Neurais (Conceito Geral):** HAYKIN, S. *Neural Networks and Learning Machines*. 3. ed. Prentice Hall, 2008.