SDUWPS: Detecção de Estresse e Esforço Físico com Sinais Fisiológicos

Anderson Cristiano Sassaki Gonçalves (Autor)
Universidade Federal de São Carlos (UFSCar)
Sorocaba, Brasil
andersoncsg@estudante.usfcar.br

Lorenzo Grippo Chiachio (Autor)
Universidade Federal de São Carlos (UFSCar)
Sorocaba, Brasil
lorenzo.chiachio@estudante.ufscar.br

***Abstract*—O monitoramento da atividade fisiológica por meio de dispositivos vestíveis apresenta-se como uma ferramenta essencial para o acompanhamento da saúde e do bem-estar em tempo real. Esses dispositivos permitem a coleta contínua de dados durante atividades cotidianas e sessões de exercício, possibilitando a identificação de padrões fisiológicos associados a estados do corpo humano, tais como relaxamento, esforço físico e estresse induzido. Este trabalho explora o uso de técnicas de Aprendizado de Máquina (Machine Learning) na análise desses sinais, visando desenvolver modelos computacionais capazes de reconhecer e classificar automaticamente esses estados fisiológicos.**

Utiliza-se uma base de dados composta por dados brutos provenientes de sensores vestíveis, incluindo frequência cardíaca, temperatura corporal, atividade eletrodérmica e acelerômetro triaxial, coletados durante sessões experimentais estruturadas de exercício e estresse induzido. A integração de sensores e algoritmos inteligentes demonstra o potencial de revolucionar a análise do desempenho físico e abrir novas perspectivas para aplicações em saúde preventiva e esportes.

***Index Terms*—Aprendizado de Máquina; Sinais Fisiológicos; Dispositivos Vestíveis; Análise de Séries Temporais.**

I. Introdução

O monitoramento contínuo de variáveis fisiológicas por meio de dispositivos vestíveis surgiu como uma ferramenta essencial para o acompanhamento da saúde e do bem-estar em tempo real. Esses dispositivos, capazes de coletar dados tanto durante atividades diárias quanto em sessões experimentais estruturadas, facilitam a identificação de padrões fisiológicos associados a distintos estados do corpo humano, como relaxamento, esforço físico (*AEROBIC*/*ANAEROBIC*) e estresse induzido (*STRESS*). O crescente interesse neste campo é impulsionado pelo potencial de utilizar dados de sensores — incluindo frequência cardíaca (HR), temperatura corporal (TEMP), atividade eletrodérmica (EDA) e acelerômetro triaxial (ACC) — para desenvolver sistemas robustos e não invasivos para avaliação proativa da saúde e avaliação de desempenho.

No entanto, um desafio fundamental no aproveitamento desses dados de sensores reside em sua natureza complexa: as medições apresentam-se como séries temporais multivariadas, frequentemente contendo ruído, *outliers* e erros de medição significativos. Dada a natureza bruta e sequencial dos sinais, um entendimento profundo da estrutura de dados subjacente é crítico antes que qualquer modelo de Aprendizado de Máquina (ML) possa ser aplicado com sucesso.

Para abordar essa questão, foi realizada uma extensa análise exploratória de dados. Essa fase inicial foi crucial não apenas para identificar problemas de qualidade nos dados, como *outliers* e valores ausentes, mas também para obter um entendimento abrangente sobre a natureza e a dinâmica dos dados que nos foram fornecidos. Essa etapa de análise e estudo da informação bruta estendeu-se para além dos sinais em si, incorporando também dados demográficos sobre os usuários que geraram as amostras. Essa abordagem sistêmica garantiu que as etapas subsequentes de engenharia de características (*feature engineering*) e treinamento do modelo fossem baseadas em uma compreensão completa tanto das respostas fisiológicas quanto do contexto do qual as amostras derivaram, levando, em última análise, a modelos de ML mais confiáveis e generalizáveis para a classificação dos estados fisiológicos.

Seguindo a motivação estabelecida pela análise inicial dos dados, este artigo pretende avaliar a eficácia de diversos algoritmos de Aprendizado de Máquina na classificação precisa de estados fisiológicos (especificamente *STRESS*, *AEROBIC* e *ANAEROBIC*) utilizando dados de séries temporais multivariadas pré-processados, provenientes de dispositivos vestíveis. Nosso objetivo principal é identificar uma abordagem de modelagem robusta — abrangendo engenharia de características, tratamento de *outliers* e arquitetura do modelo — que proporcione um desempenho ideal e uma linha de base confiável para pesquisas futuras focadas na detecção de estados fisiológicos em tempo real em aplicações clínicas e esportivas.

O restante deste artigo está organizado da seguinte forma: a Seção II detalha a análise exploratória de dados realizada. A Seção III aborda o pré-processamento dos dados, as técnicas de engenharia de *features* aplicadas aos dados de séries temporais, incluindo métodos para tratamento de *outliers* e também a aplicação dos modelos de aprendizado de máquina. A Seção IV apresenta uma análise abrangente e a discussão dos resultados experimentais e do desempenho dos modelos e, finalmente, a Seção V fornece a conclusão deste projeto.

II. Análise Exploratória

Nessa etapa do projeto, o foco foi entender os dados brutos e analisar suas possíveis falhas, como possíveis *outliers*, informação faltando e dados mal coletados. Para atingir esse objetivo, seguimos uma série de etapas, gerando vários gráficos acerca das métricas que nos foram fornecidas.

Primeiro, observamos a distribuição das classes dentro das amostras de treino por meio de um gráfico de barras (Figura 1).

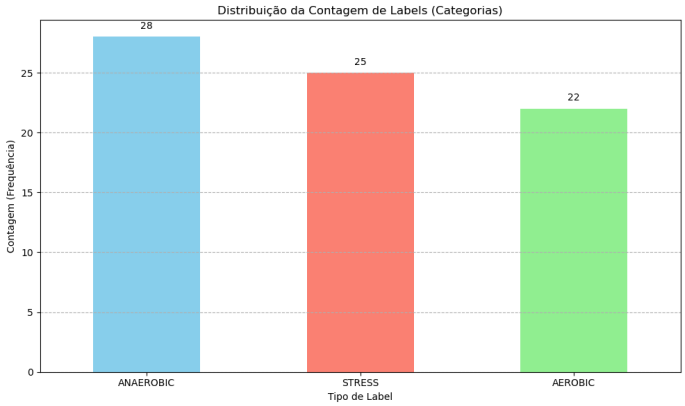


Figure 1: Distribuição de classes dentro das amostras de treino.

Como podemos verificar no gráfico, a contagem de amostras por categoria aproxima-se de uma distribuição uniforme, sendo 28 amostras rotuladas como *ANAEROBIC*, 25 como *STRESS* e 22 como *AEROBIC*. Essa homogeneidade no conjunto de treinamento é uma característica desejável.

Seguindo o *pipeline* da análise exploratória, foi feita uma verificação da integridade dos dados brutos, visando identificar arquivos ausentes ou corrompidos nos diretórios das amostras. Durante essa varredura, a estrutura de arquivos da maioria dos participantes mostrou-se consistente. Contudo, detectou-se uma anomalia específica na amostra referente ao usuário U_89740, pois o arquivo IBI.csv se encontrava vazio. Essa ausência de dados foi mantida, mas sem descartar a amostra como um todo, sendo que no pré-processamento e na seleção de *features* foram considerados apenas os dados dos outros sensores.

Dando continuidade à exploração dos dados, seguimos para a caracterização dos sinais capturados pelos dispositivos vestíveis. O conjunto de dados abrange seis modalidades sensoriais distintas, e, para uma melhor compreensão destes, nós desenhamos os gráficos de uma amostra aleatória, chegando a um melhor entendimento geral sobre o que cada sensor visa captar. Nesse sentido, ressaltamos que para a utilização do acelerômetro triaxial, tanto no pré-processamento quanto na seleção de *features*, foi feita uma conversão para magnitude do vetor de sinal, calculada a partir dos três parâmetros da medida do ACC ($\sqrt{X^2 + Y^2 + Z^2}$). Essa abordagem transforma os três sinais direcionais em uma única métrica que representa a intensidade total do movimento. Isso simplifica a visualização e cria uma característica mais robusta e intuitiva.

Dessa forma, decidimos obter uma visão geral da distribuição dos dados dentro de cada classe, o que também aumentou o nosso entendimento das métricas e começou a nos dar uma visão do *pipeline* de pré-processamento que teríamos que seguir. Realizamos essa distribuição desenhando gráficos de linha e de *boxplot* para visualizar os *outliers* e outras medidas descritivas de maneira mais clara. Nas figuras 2 e 3 podemos ver o desenho das distribuições dos dados da temperatura de algumas amostras de treino pertencentes à classe *AEROBIC*.

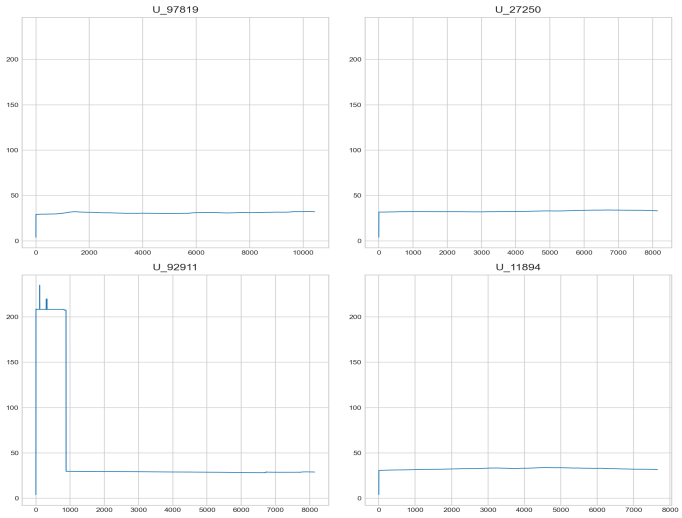


Figure 2: Distribuição dos dados das amostras de temperatura dentro da classe *AEROBIC* (algumas amostras)

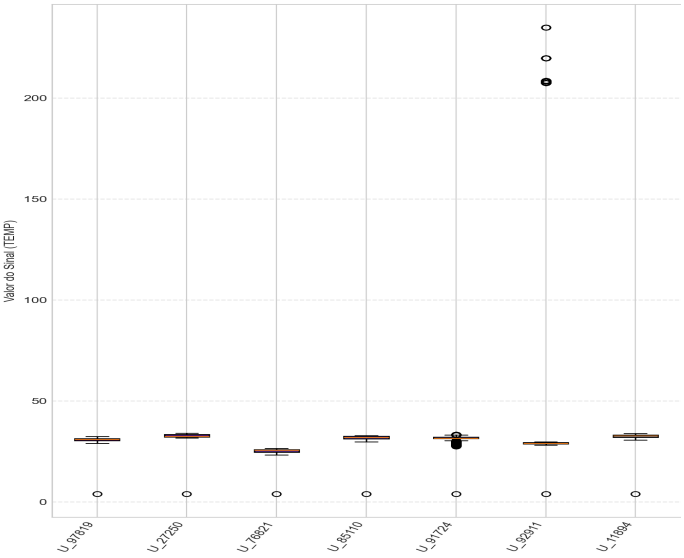


Figure 3: Distribuição dos dados das amostras de temperatura dentro da classe *AEROBIC* utilizando o gráfico de *boxplot* (algumas amostras)

A análise exploratória contemplou também o estudo dos dados demográficos dos participantes, abrangendo atributos como idade, gênero, peso e altura.

III. Pré-processamento e Aplicação dos Modelos

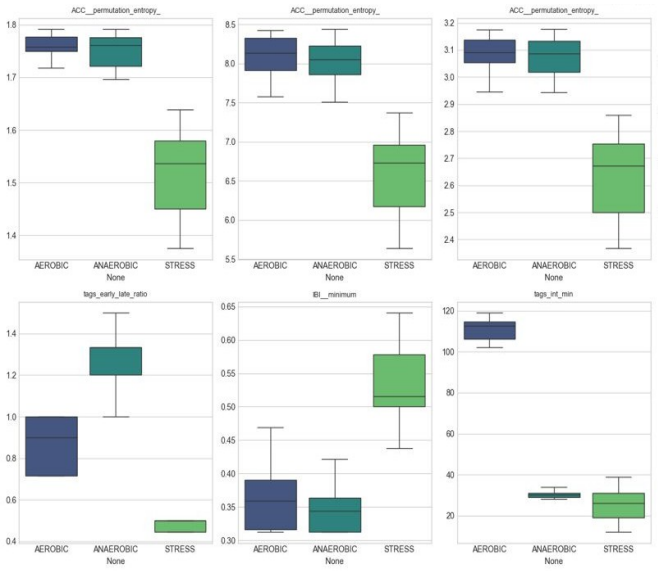
A. Pré-Processamento

Para o pré-processamento foram feitas algumas tentativas envolvendo técnicas diferentes que, por fim, geravam resulta-

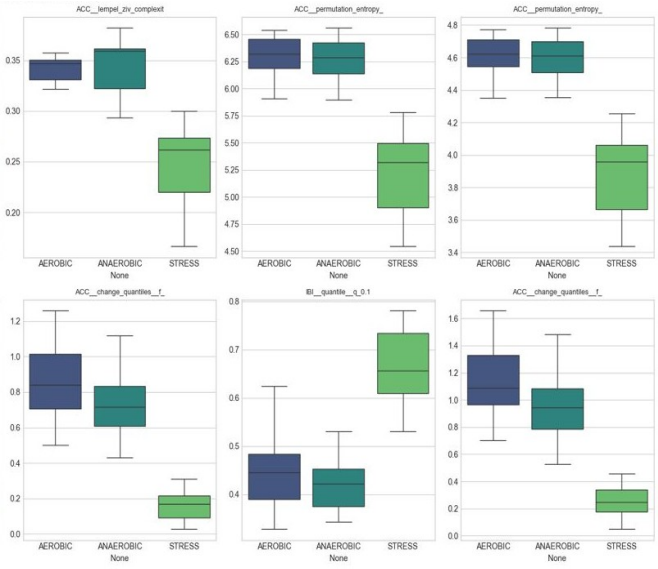
dos ruins durante a análise dos modelos. Uma das estratégias investigadas consistiu na substituição dos *outliers* pela média dos valores do sensor, calculada especificamente para a classe à qual a amostra pertence. Além dessa, outra tentativa falha foi a técnica de janelamento com as amostras, a qual envolvia a segmentação das séries temporais para uma melhor utilização das nuances dos dados coletados pelos sensores no aprendizado dos modelos.

Assim, o pipeline de pré-processamento adotado neste projeto inicialmente trata os *outliers* com a técnica do *hard clipping*, onde valores ruidosos foram substituídos utilizando limites estatísticos definidos pelo dobro do intervalo interquartil. Após a remoção dos *outliers*, os dados seguem para serem normalizados para garantir a consistência escalar entre os diferentes sensores durante o treinamento. Concomitantemente, o pré-processamento dos dados demográficos também foi realizado, preenchendo os dados categóricos foram preenchidos com a moda e os não-categóricos, com a média dos outros valores do conjunto fornecido. A etapa seguinte de engenharia de atributos utilizou a biblioteca `tsfresh` para a extração e seleção automática de *features* significativas das séries temporais, essa biblioteca atua automatizando a extração de um número grande de características matemáticas das séries temporais, abrangendo desde estatísticas descritivas simples até métricas mais complexas.

Por fim, a qualidade do espaço de atributos gerado foram avaliadas por meio de técnicas de redução de dimensionalidade (PCA e t-SNE) e da inspeção visual da distribuição das 12 melhores *features* selecionadas através de gráficos de *boxplot*, como pode ser visto em 4a e em 4b. A técnica de t-SNE (*t-Distributed Stochastic Neighbor Embedding*) é uma técnica de redução de dimensionalidade não linear ideal para visualização de dados de alta complexidade. Diferente do PCA, que foca na variância global, o t-SNE prioriza a preservação da estrutura local dos dados, aproximando no espaço bidimensional pontos que são vizinhos no espaço original. Com essa técnica, ganhamos uma perspectiva a mais sobre os dados redimensionados, além da do PCA.



(a) Primeiras 6 features



(b) Últimas 6 features

Figure 4: As 12 *features* mais relevantes extraídas pelo *tsfresh*

B. Aplicação dos modelos de Aprendizado de Máquina

Para gerar os modelos finais que iriam classificar os estados fisiológicos das amostras, utilizamos bibliotecas amplamente reconhecidas na literatura de Aprendizado de Máquina por sua robustez e eficiência computacional. O *framework* `scikit-learn` serviu como base estrutural para a implementação dos *pipelines* de avaliação e dos algoritmos clássicos, incluindo *K-Nearest Neighbors* (KNN), *Support Vector Machines* (SVM), *Naive Bayes*, Regressão Logística e *Multi-layer Perceptron* (MLP). Adicionalmente, visando explorar o desempenho de técnicas avançadas de *ensemble*, foram integradas as bibliotecas especializadas `XGBoost` e `LightGBM`, que oferecem implementações otimizadas de *Gradient Boosting*. Essa abordagem nos permitiu testar diferentes modelos e então, na etapa de análise, escolher o melhor para o problema.

Para o nosso projeto a escolha de incluir o *Random Forest* justifica-se pela robustez do modelo através da técnica de *Bagging*, a qual reduz a variância do modelo e oferece estabilidade

frente a dados com maior número de *outliers*. Em adição, a incorporação do *XGBoost* e do *LightGBM* visou explorar o potencial do *Gradient Boosting*. Estes algoritmos destacam-se pela otimização sequencial dos erros e pela implementação de regularizações avançadas. A utilização conjunta dessas três abordagens permitiu não apenas buscar a maximização das métricas de desempenho, mas também comparar a eficácia da agregação paralela de árvores por meio do *Random Forest* contra a construção sequencial e corretiva *Boosting* na classificação dos estados fisiológicos.

Por conseguinte, para maximizar o desempenho preditivo adotou-se uma estratégia experimental de variação de hiperparâmetros, onde diferentes arquiteturas e parâmetros de regularização foram testados sistematicamente. Esse processo de ajuste fino permitiu identificar a configuração ótima para cada algoritmo, assegurando que a comparação de desempenho refletisse o potencial máximo de cada classificador no contexto da detecção de estados fisiológicos.

IV. Análise dos Resultados

A avaliação comparativa dos modelos foi conduzida por meio de um *pipeline* que foi aplicado a todos os algoritmos testados. Para cada classificador, gerou-se a matriz de confusão, permitindo uma visualização detalhada dos erros de classificação cruzada entre as classes. Complementarmente, utilizou-se a métrica AUC (*Area Under the Curve*), derivada da curva ROC, como indicador principal de robustez. Essa métrica, que é obtida pela área da curva do ROC (Receiver Operating Characteristic), que por sua vez é a medida de todos os limites de classificação (*thresholds*) possíveis do desempenho de um modelo de classificação binário, foi fundamental para quantificar a capacidade de generalização dos modelos, servindo de principal guia para a escolha da arquitetura definitiva que melhor soluciona o problema.

A análise dessas métricas orientou o refinamento final da estratégia de modelagem. Assim, observamos que os experimentos demonstraram que as métricas provenientes do arquivo de marcação (`tags.csv`) foram determinantes para a qualidade da classificação. Tal fato também pôde ser observado a partir dos 12 gráficos boxplot que geramos para as 12 *features* mais importantes extraídas pelo `tsfresh`.

V. Conclusão

Para a conclusão desse projeto, seguindo todo o *pipeline* de processamento, desde a análise dos dados até a aplicação do modelo e a análise do desempenho dos mesmos, decidimos que o modelo de *XGBoost* é o melhor para o problema apresentado, obtendo uma AUC de 100% em cima do conjunto de treino e, aproximadamente, 99.6% no conjunto de teste. Além disso, esse modelo obteve f1-scores de 95% na classe *AEROBIC*, 89% na classe *ANAEROBIC* e 100% na classe *STRESS*. Por fim, justificando a nossa escolha obtivemos a matriz de confusão e a curva ROC (*Receiver Operating Characteristic*) que podem ser visualizadas em 5 e 6, respectivamente.

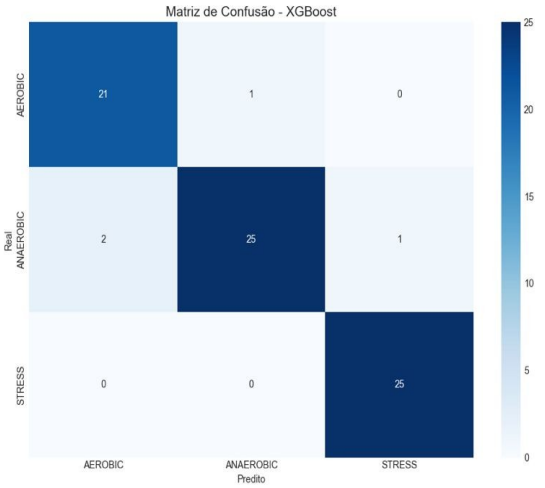


Figure 5: Matriz de confusão para o *XGBoost*

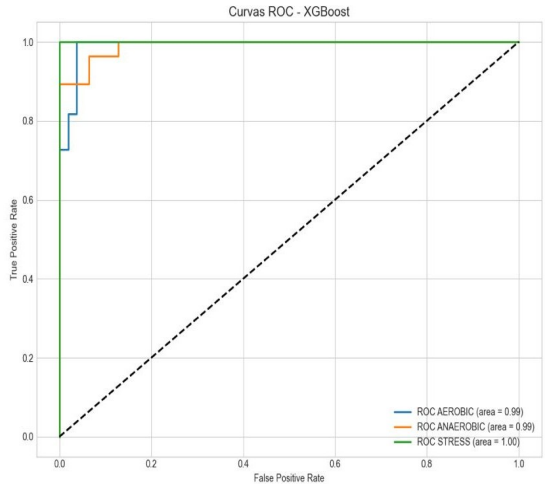


Figure 6: Curva ROC para o *XGBoost*

References

[1] Q. Wang, Y. Zhang and H. Liu, "A Classification Method for Electricity Users Based on the LightGBM Algorithm," in *Proceedings of the 2023 IEEE Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 2023, pp.—, doi:10.1109/ICEEICT56924.2023.10157812. ([journal.esrgroups.org][1])

[2] "Configure XGBoost classification/regression solutions," in *ServiceNow Documentation – Zurich Intelligent Experiences*, Serviço da ServiceNow. Acessado em 27 nov. 2025.

[3] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," *Expert Systems with Applications*, vol. 237, 2024, art. no. 121549, doi:10.1016/j.eswa.2023.121549. ([sciencedirect.com][1])

[4] "Introdução ao t-SNE: Redução de dimensionalidade não linear e visualização de dados," *DataCamp*, 24 abr. 2024. Acessado em 27 nov. 2025. ([datacamp.com][1])

[5] L. H. Barbosa Filho, "Como criar janelas móveis de séries temporais usando o Python," *Análise Macro – Mercado Financeiro*, publicado em 31 mar. 2025. Acessado em 27 nov. 2025. ([analisemacro.com.br][1])

[6] "AUC e a curva ROC no aprendizado de máquina," *DataCamp*, autor Vidhi Chugh. Acessado em 27 nov. 2025. ([datacamp.com][1])