Conference Paper Title*

1st Anderson Nascimento
Departamento de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
andersonmoura.cn@gmail.com

2rd J. Rodrigo Nascimento
Departamento de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
rodri.nasc@alu.ufc.br

3nd Osvaldo Cavalcante Neto
Departamento de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
osvaldomedeiros@alu.ufc.br

4th Vinícius Lira
Departamento de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brasil
viniciuslavorlira@gmail.com

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introdução

A sensação de estresse é definida como a reação natural do organismo humano a situações de perigo, ameaça ou medo. Essa ferramenta do corpo é comum a todo humano e nos coloca em estado de alerta, provocando sintomas físicos ou emocionais, como dor de cabeça, batimentos cardíacos acelerados, sensação de desgaste fisíco, etc. Além disso, o estresse e seus sintomas são comumente atrelados ao desenvolvimento de doenças cardiovasculares e mentais, o que mostra a importância de encontrarmos formas de prever ou detectar esse tipo de reação do corpo.

Na literatura, a relação entre o estresse e alguns parâmetros fisiologicos atrelados ao HRV (Heart Rate Variability) já tem sido estudada, indicando que as varíaveis HRV são afetadas diretamente pelo estresse [**?**], com atividades que geram estresse ocupacional impactando diretamente na redução da variação de frequência cardíaca [**?**], o que leva a menor flexibilidade do sistema nervoso autônomo e maior predominância da atividade simpática, o que gera o sentimento de fadiga física e mental.

Atualmente, com o avanço das tecnologias de monitoramento cardíaco [**?**], como sensores de baixo custo e "wearables" e o aumento na disponibilização de bases de dados reais colhidos de pacientes, torna-se possível analisar e visualizar o comportamento do estresse e suas consequencias. A análise desses dados nos permite identificar padrões fisiológicos e compreender sinais característicos da sensação de estresse, nos permitindo a criação de ferramentas baseadas em aprendizado estatístico para detecção e prevenção de estresse, capazes de auxiliar na detecção precoce e na prevenção do estresse em contextos clínicos, ocupacionais e cotidianos.

Nesse cenário, a análise exploratória de dados (EDA) surge como uma ferramenta essencial para compreender os dados, avaliar sua qualidade e identificar relações iniciais entre variáveis fisiológicas e níveis de estresse. O dataset Heart Rate Prediction to Monitor Stress Level, fornecido pela plataforma de dados Kaggle surge nesse contexto como uma fonte de dados de grande valor para investigação científica.

Portanto, esse artigo tem como intuito principal realizar uma análise exploratória de dados sobre o referido dataset, buscando compreender sua estrutura, avaliar a qualidade das variáveis e identificar possíveis correlações entre parâmetros fisiológicos e níveis de estresse.

II. Metodologia

A. Coleta de Dados

O presente estudo utilizou o conjunto de dados "*Heart Rate Prediction to Monitor Stress Level*" (Shanawad, 2021), disponível na plataforma Kaggle. O dataset reúne atributos fisiológicos derivados de sinais de eletrocardiograma (ECG) obtidos em diferentes indivíduos sob distintas condições de estresse, com o objetivo de analisar a relação entre variabilidade da frequência cardíaca e níveis de estresse.

Os dados foram fornecidos em seis arquivos .csv, divididos em subconjuntos de treinamento e teste, organizados nos domínios do tempo, frequência e não linearidade, contendo variáveis como MEAN_RR, RMSSD, LF/HF, SD1, Sampen, entre outras. O conjunto de treinamento, utilizado neste trabalho, apresenta $N = 369.289$ amostras e $D = 36$ variáveis preditoras, além de duas variáveis-alvo: HR (frequência cardíaca) e condition (nível de estresse).

A variável condition possui L = 3 classes, distribuídas em no stress ($54,18\%$ das amostras), interruption ($28,47\%$ das amostras) e time pressure ($17,35\%$ das amostras), indicando leve desbalanceamento entre categorias. A análise descritiva inicial foi conduzida em `Python` com a biblioteca Pandas, permitindo identificar dimensões, distribuição de classes e consistência dos dados. Apenas o conjunto de treinamento foi utilizado, visto que os arquivos de teste não contêm a variável HR, inviabilizando validação supervisionada. Dessa forma, todas as análises exploratórias e procedimentos de pré-processamento foram conduzidos exclusivamente sobre o

conjunto de treinamento, abrangendo os três domínios de características.

B. Análise Exploratória dos Dados

O objetivo principal da análise exploratória é compreender a estrutura interna dos dados e identificar as variáveis mais pertinentes, ou seja, quais características fisiológicas possuem maior poder discriminatório para diferenciar os três estados de estresse definidos no estudo: no stress, interruption e time pressure. Além disso, por meio da análise, é possível detectar eventuais problemas, como a redundância, antes de se proceder à modelagem. Esta etapa é crucial para garantir a robustez das etapas futuras.

Visualização dos Dados

Para compreender a distribuição dos preditores, cada variável foi analisada individualmente, com suas estatísticas descritivas fundamentais: a média, o desvio padrão e a assimetria calculadas. As Tabelas **??**, **??** e **??** apresentam essas estatísticas para cada preditor do conjunto de dados, permitindo identificar a magnitude e a dispersão de cada variável fisiológica. A análise foi realizada inicialmente de maneira incondicional, ou seja, sobre o conjunto de dados completo e, posteriormente, de modo condicional, levando em consideração as classes, visando avaliar o poder discriminatório de cada variável. Além disso, para cada uma das análises, foi utilizado histogramas para visualizar a forma de cada distribuição e boxplots para identificar outliers. A Figura **??** mostra o histograma e o box-plot incondicional da variável HR, que apresenta uma leve assimetria positiva e uma maior concentração dos dados entre 65-85 bpm.

A Figura **??** apresenta os histogramas por classe de MEAN_RR, que mostra claramente ser uma variável discriminatória, pois, dependendo da classe, temos diferentes distribuições.

A análise bivariada investigou as relações entre os pares de preditores utilizando diagramas de dispersão, com pontos coloridos conforme a classe *condition*, permitindo avaliar visualmente a separação entre classes. Complementarmente, foi calculada uma matriz de correlação e visualizada como mapa de calor, facilitando a identificação de padrões de dependência linear e multicolinearidade entre os preditores.

C. Refinação dos dados

A fim de otimizar o conjunto de dados para a análise, foi realizada uma etapa rigorosa de seleção de características para remover preditores redundantes ou pouco informativos, utilizando dois critérios principais.

Para mitigar o problema de multicolinearidade entre os preditores, foi adotado um método heurístico baseado na matriz de correlação de Pearson, conforme proposto por Kuhn e Johnson [**?**]. O algoritmo remove o número mínimo de variáveis necessárias para que todas as correlações pareadas fiquem abaixo de um limiar, que foi definido para esse estudo como 0.8. Em cada iteração, identifica-se o par mais correlacionado, calcula-se a média das correlações de cada variável e sugere-se a remoção daquela com maior média. O processo é repetido até que não restem correlações acima do limiar, reduzindo redundâncias e melhorando a estabilidade dos modelos preditivos. Para esse estudo, a variável HR não foi considerada na aplicação do algoritmo, devido ao seu uso como variável-alvo no futuro.

TABLE I
ESTATÍSTICAS DESCRITIVAS DOS PREDITORES (PARTE 1 DE 3).

Estatística	MEAN_RR	MEDIAN_RR	SDRR	RMSSD	SDSD	SDRR_RMSSD	HR	pNN25	pNN50	KURT	SKEW
Média	846.65	841.97	109.35	14.98	14.98	7.40	73.94	9.84	0.87	0.52	0.04
Desvio-padrão	124.60	132.32	77.12	4.12	4.12	5.14	10.34	8.20	0.99	1.79	0.70
Assimetria	0.65	0.93	2.36	0.40	0.40	3.71	0.41	1.20	1.26	5.72	1.22

TABLE II
ESTATÍSTICAS DESCRITIVAS DOS PREDITORES (PARTE 2 DE 3).

Estatística	MEAN _REL_RR	MEDIAN _REL_RR	SDRR _REL_RR	RMSSD _REL_RR	SDSD _REL_RR	SDRR_RMSSD _REL_RR	KURT _REL_RR	SKEW _REL_RR	VLF	VLF_PCT	LF
Média	0.00	0.00	0.02	0.01	0.01	2.01	0.52	0.04	2199.58	64.29	946.53
Desvio-padrão	0.00	0.00	0.01	0.00	0.00	0.38	1.79	0.70	1815.77	16.77	574.17
Assimetria	0.11	−0.95	0.87	1.26	1.26	0.84	5.72	1.22	1.96	−0.41	1.35

TABLE III
ESTATÍSTICAS DESCRITIVAS DOS PREDITORES (PARTE 3 DE 3).

Estatística	LF_PCT	LF_NU	HF	HF_PCT	HF_NU	TP	LF_HF	HF_LF	SD1	SD2	sampen	higuci	datasetId
Média	34.10	95.57	39.25	1.62	4.43	3185.36	115.98	0.05	10.59	154.18	2.06	1.18	2.00
Desvio-padrão	16.04	4.12	45.40	1.76	4.12	1923.23	360.86	0.05	2.91	109.17	0.21	0.06	0.00
Assimetria	0.43	−1.65	2.48	2.02	1.65	1.45	9.78	2.16	0.40	2.36	−3.09	0.34	0.00

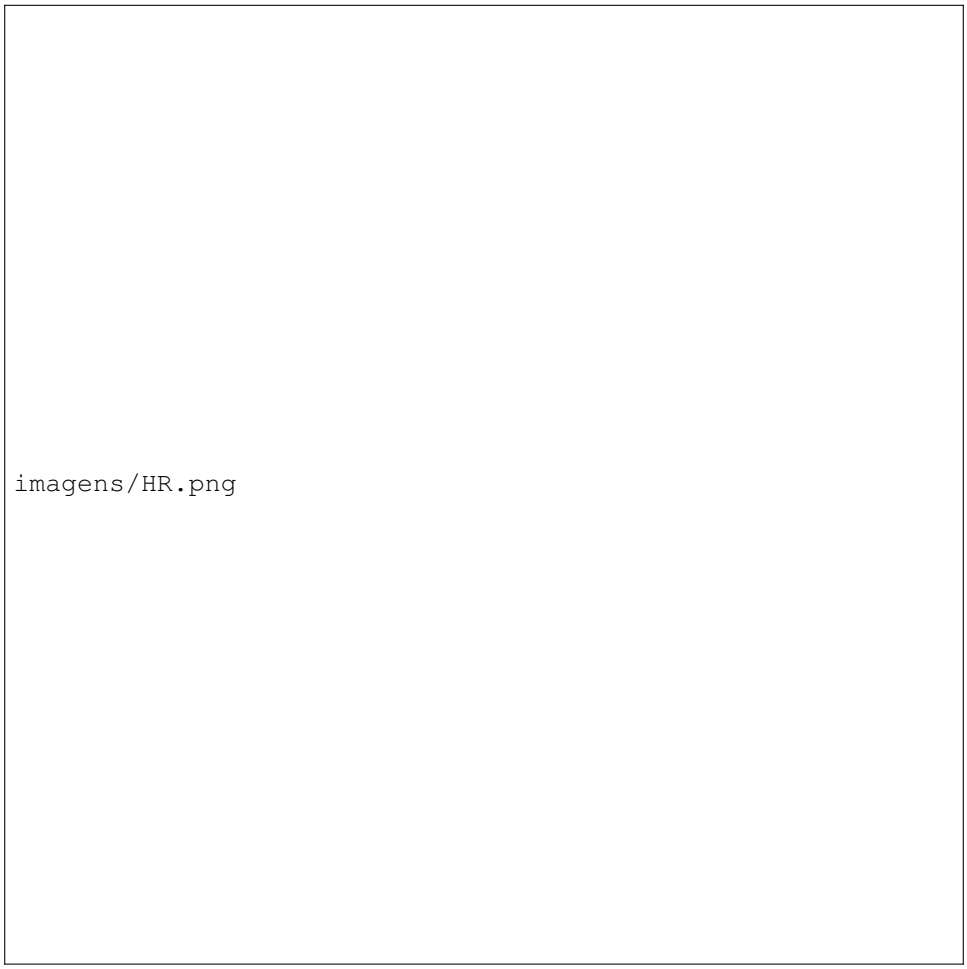

Fig. 1. Histograma e box-plot incondicional de HR.

Em segundo lugar, foi realizada uma análise discriminante baseada nos resultados da análise univariada condicional. As variáveis que demonstraram baixo poder de separação entre as classes de estresse, ou seja, cujas distribuições se sobrepunham significativamente entre as diferentes condições, foram consideradas não informativas e também foram eliminadas.

Este processo de dupla filtragem resultou em um conjunto final e otimizado de preditores, que foi então utilizado para a subsequente Análise de Componentes Principais.

D. Análise de Componentes Principais

A Análise de Componentes Principais (PCA) é uma técnica estatística multivariada utilizada para a redução de dimensionalidade e extração de características. Sua principal importância reside na capacidade de transformar um conjunto complexo de variáveis correlacionadas em um novo conjunto, dimensionalmente menor, de variáveis não correlacionadas, chamadas componentes principais, retendo o máximo possível da variância original dos dados. A metodologia para a sua aplicação neste trabalho seguiu os seguintes passos propostos por Kuhn e Johnson [**?**]:

1) **Padronização dos Dados:** O processo iniciou-se com centralização na média, onde a média de cada variável é subtraída de seus valores, e o escalonamento pela variância, onde o resultado é dividido pelo desvio padrão, de forma a assegurar que a análise se concentrasse na estrutura de variância em detrimento da localização espacial dos dados.
2) **Cálculo da Matriz de Covariância e Decomposição em Autovalores e Autovetores:** Com os dados padronizados, foi novamente calculada uma matriz de covariância. Ela foi então decomposta para encontrar seus autovalores (λ) e autovetores ($\vec{v}$). Nesta etapa, os autovetores representam as direções dos componentes principais, enquanto os autovalores correspondentes quantificam a magnitude da variância capturada por cada autovetor.
3) **Projeção:** Os autovetores foram, então, ordenados de forma decrescente com base em seus autovalores, com os dados sendo projetados nos dois componentes principais com os maiores autovalores (PC1 e PC2).

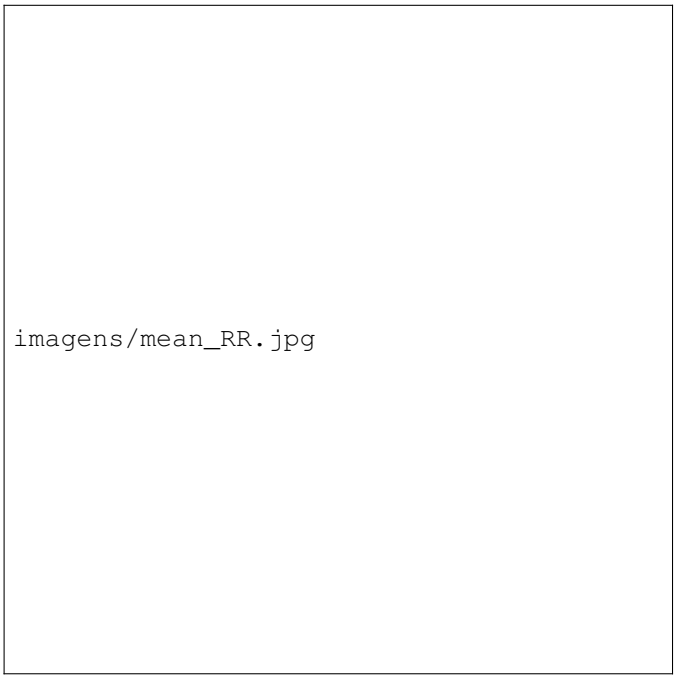

Fig. 2. Histogramas de MEAN_RR por classe.

III. Resultados

A. Estrutura dos dados e checagens iniciais

Durante o preprocessamento, verificamos que os dados não continham valores ausentes, mas uma presença considerável de outliers. Os intervalos e valores observados para as variáveis estavam dentro do esperado fisiologicamente. A frequência cardíaca variou aproximadamente dentro da faixa adulta típica (cerca de 50–180 bpm entre os indivíduos), com médias em torno de valores normais em repouso ($\approx$ 60-80 bpm) e elevações compatíveis com situações de maior demanda experimental.

A presença do atributo condition permitiu comparar os grupos, mas as médias não seguiram o padrão fisiológico clássico (HR maior e HRV menor em estresse), indicando que as classes representam indivíduos diferentes e não medições repetidas. Ainda assim, a Figura **??** evidencia o comportamento global das variáveis entre condições.

B. Achados principais

Após o pré-processamento e análise exploratória dos dados, identificamos padrões que relacionam as métricas de variabilidade da frequência cardíaca (HRV) com a frequência cardíaca medida (HR).

1) Relação entre MEAN_RR, MEDIAN_RR e HR: Inicialmente, há uma correlação inversa entre o intervalo R-R médio (MEAN_RR) e a frequência cardíaca (HR), indicando que intervalos menores entre batimentos estão associados a HR mais alta. O que é esperado, pois quanto menor o RR, maior o número de batimentos por minuto [**?**].

Em repouso o RR costuma ser $\approx$ 1 segundo (HR $\approx$ 60 bpm), enquanto RR de 0,5 s implica HR $\approx$ 120 bpm (taquicardia). Em que, considera-se taquicardia um ritmo cardíaco de repouso acima de 100 bpm [**?**], valor este que de fato foi ultrapassado em alguns registros do conjunto de dados, sugerindo episódios de estresse ou esforço significativos.

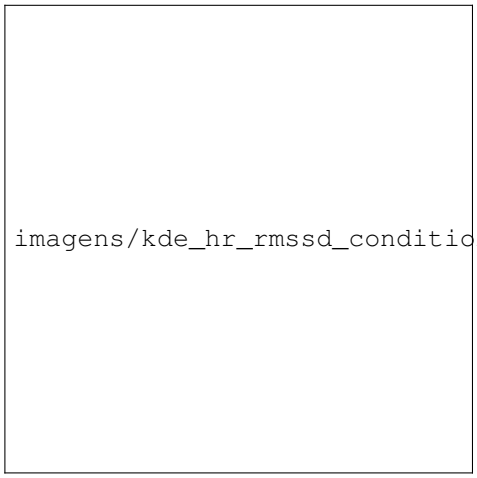

Fig. 3. Densidade bivariada HR × RMSSD nas diferentes condições.

Fig. 4. Dispersão HR × MEAN_RR por classe

A mesma correlação forte e negativa ocorre entre MEDIAN_RR e HR ($r \approx -0,93$). O que é esperado por definição, já que MEDIAN_RR e MEAN_RR carregam essencialmente a mesma informação fisiológica da frequência cardíaca, sendo altamente correlacionados.

2) Variabilidade de curto e longo prazo: Observamos que métricas clássicas de variabilidade, especialmente aquelas dominadas pelo tônus parassimpático, tendem a diminuir quando a frequência cardíaca está elevada. Por exemplo, indicadores como RMSSD (raiz quadrada da média dos quadrados das diferenças sucessivas de RR) e pNN50 (porcentagem de intervalos RR sucessivos que diferem em mais de 50ms) exibiram valores menores em casos de HR alto. O que sugerem menor variabilidade batimento a batimento durante episódios de frequência elevada, coincidindo com o esperado

Fig. 5. Dispersão MEAN_RR × MEDIAN_RR por classe

em situações de estresse ou ativação simpática.

Na literatura, condições de estresse psicológico costumam se associar a HRV reduzida, incluindo queda em métricas como pNN50 e potência de alta frequência (HF), indicando menor influência vagal, ao mesmo tempo em que a razão LF/HF se eleva [**?**]. Nossos dados seguem esse padrão geral: as condições associadas a maior demanda apresentaram pNN50 e RMSSD mais baixos e LF/HF mais altos, refletindo um desequilíbrio autonômico compatível com a ativação simpática. Essa combinação de menor variabilidade de curto prazo e aumento relativo das oscilações de longo prazo sugere menor flexibilidade autonômica sob estresse, em consonância com o descrito na literatura [**?**].

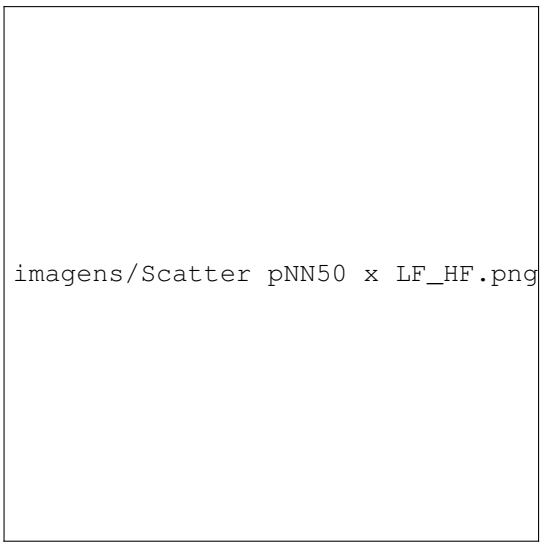

Fig. 6. Dispersão pNN50 × LF/HF por condição.

3) Espectro e razão: No domínio da frequência, vimos que, à medida que aumentava a demanda das tarefas (no stress → interruption → time pressure), a potência na banda de alta frequência (HF, 0,15–0,4 Hz) tendia a diminuir, enquanto a potência de baixa frequência (LF, 0,04–0,15 Hz) mostrava aumento relativo, resultando em elevação da razão LF/HF (Figura **??**). Esse comportamento indica uma redução da modulação vagal e maior predominância simpática, conforme descrito em estudos sobre respostas autonômicas ao estresse [**?**].

As três condições do conjunto de dados apresentaram essa tendência de aumento progressivo da razão LF/HF, embora com grande dispersão.

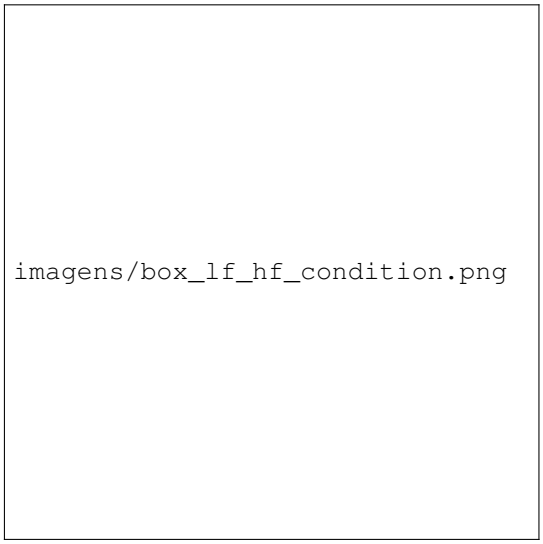

Fig. 7. Boxplot LF/HF por condição.

Quando a razão inversa (HF/LF) é observada, a interpretação se mantém invertida: em situações de maior demanda (menor influência vagal), HF diminui e LF aumenta, levando a LF/HF mais altos e, portanto, HF/LF mais baixos.

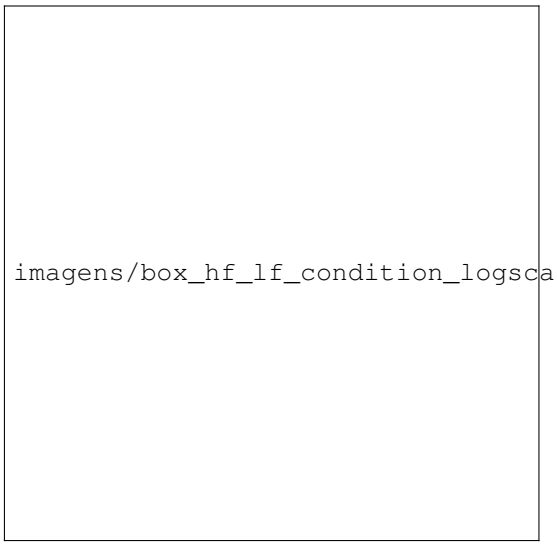

Fig. 8. Boxplot HF/LF por condição.

É importante lembrar que, embora HF represente predominantemente a atividade parassimpática e LF inclua componentes simpáticos e barorreflexos, a razão LF/HF deve ser interpretada com cautela, servindo apenas como indicador relativo de deslocamento do balanço autonômico.

4.4 Potência total e distribuição espectral: Entre os índices de potência, a variável *TP* (total power) representa a energia global da variabilidade dos intervalos RR, enquanto *VLF_PCT* descreve a contribuição relativa das oscilações de muito baixa frequência [**?**].

Na Figura **??** observa-se que a potência total (*TP*) apresentou valores ligeiramente mais altos sob *time pressure*. Embora a literatura clássica reporte reduções de *TP* e *HF* em estresse agudo [**?**], este comportamento divergente pode indicar particularidades do conjunto de dados e da segmentação dos sinais, não necessariamente uma inconsistência fisiológica.

Assim como, *VLF_PCT* mostrou valores ligeiramente mais altos nessa condição, sugerindo maior participação relativa dos componentes de muito baixa frequência.

Essas variações, embora sutis, indicam que a resposta ao aumento de demanda não se limitou à redistribuição entre bandas (LF e HF), mas também envolveu alterações na potência total do espectro.

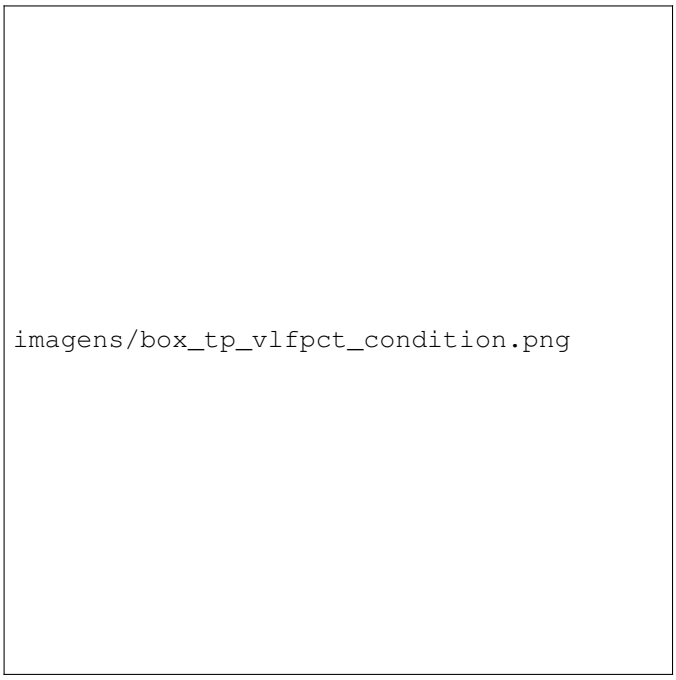

Fig. 9. Boxplots por condição: *TP*, *VLF_PCT*

C. Redundância e seleção

Vários dos preditores calculados apresentam definições semelhantes ou alto grau de correlação mútua. Por exemplo, RMSSD e pNN50 ambos quantificam variabilidade de curto prazo dominada pelo sistema parassimpático (vagal) e, como previsto, mostraram-se fortemente correlacionados entre si nos dados ($r \approx 0.79$). O que já era conhecido, tanto que RMSSD é frequentemente preferido ao pNN50 por ter propriedades estatísticas mais estáveis [**?**].

Essa redundância sugere que nem todos os preditores fornecem informação inédita para o modelo de regressão ou de classificação. Sabendo disso e das informações anteriores, escolhemos o subconjunto de preditores formado por

(i) **Cronotropia** (*HR*, *MEDIAN_RR*);

(ii) **Variabilidade de curto prazo vagal** (*RMSSD*, *SDSD_REL_RR*);

(iii) **Dispersão global de longo prazo** (*SD2*, *SDRR_REL_RR*);

(iv) **Componentes espectrais** (*LF*, *HF*, *VLF_PCT*, *TP*); e

(v) a **Razão espectral inversa** (*HF_LF*).

Métricas não lineares, como a entropia amostral (Sampen), foram calculadas mas não mostraram poder discriminativo relevante.

D. Análise de Componentes Principais (PCA) e Biplot

O PCA aplicado aos 12 preditores padronizados explicou cerca de **71,2%** da variância total nas duas primeiras componentes (*PC1* = 40,5%, *PC2* = 30,7%). O mapa *PC1*×*PC2* exibiu sobreposição entre classes, mas com tendência ordenada: registros de *time pressure* concentraram-se na região associada a *HR* mais alta e *MEDIAN_RR* menor, enquanto *no stress* ocupou o quadrante oposto, coerente com maior modulação vagal.

A *PC1* representou o eixo de controle cardíaco (com *HR* e *MEDIAN_RR*), enquanto a *PC2* refletiu diferenças de potência espectral (*HF*, *LF*, *TP*). No *biplot* (Figura **??**), as cargas (*loadings*) reforçam essa separação: *HR* e *MEDIAN_RR* mostram orientação oposta, confirmando correlação negativa, e as variáveis *HF* e *HF_LF* agrupam-se, indicando captação conjunta do componente vagal.

Em síntese, o *biplot* evidencia um eixo principal associado à frequência cardíaca e outro à variação espectral, reproduzindo o padrão fisiológico esperado e mostrando que as condições se distribuem de forma contínua ao longo do espectro autonômico.

Fig. 10. Biplot das duas primeiras componentes principais (PC1 e PC2) obtidas a partir dos 12 preditores padronizados.

IV. Conclusão

Em conclusão, a explicação e discussão crítica dos resultados do pré-processamento reforçam que os dados estão consistentes com os princípios da fisiologia cardíaca. O pré-processamento, incluindo padronização e eliminação de variáveis altamente correlacionadas, preservou a diversidade fisiológica entre métricas cronotrópicas, temporais e espectrais. Dessa forma, conseguimos enriquecer a análise exploratória com contexto biológico, oferecendo ao leitor uma compreensão mais profunda de como os resultados do pré-processamento se conectam à realidade fisiológica subjacente aos dados.

References

[1] Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH. Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature. Psychiatry Investig. 2018 Mar;15(3):235-245. doi: 10.30773/pi.2017.08.17. Epub 2018 Feb 28. PMID: 29486547; PMCID: PMC5900369.

[2] Järvelin-Pasanen S, Sinikallio S, Tarvainen MP. Heart rate variability and occupational stress-systematic review. Ind Health. 2018 Nov 21;56(6):500-511. doi: 10.2486/indhealth.2017-0190. Epub 2018 Jun 16. PMID: 29910218; PMCID: PMC6258751.

[3] Salai, Mario, Vassányi, István, Kósa, István, Stress Detection Using Low Cost Heart Rate Sensors, Journal of Healthcare Engineering, 2016, 5136705, 13 pages, 2016. https://doi.org/10.1155/2016/5136705

[4] R. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," Frontiers in Public Health, vol. 5, p. 258, Apr. 2017. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5900369/

[5] American Heart Association, "Tachycardia (Fast Heart Rate)," Heart.org, 2024. [Online]. Available: https://www.heart.org/en/health-topics/arrhythmia/about-arrhythmia/tachycardia–fast-heart-rate.

[6] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation and clinical use," Circulation, vol. 93, no. 5, pp. 1043–1065, 1996. [Online]. Available: https://www.escardio.org/static-file/Escardio/Guidelines/Scientific-Statements/guidelines-Heart-Rate-Variability-FT-1996.pdf

[7] Kuhn, M., & Johnson, K. (2018). Applied predictive modeling. Springer.