Análise Exploratória de Dados da Frequência Cardíaca Para Classificação de Níveis de Estresse

Vitor Mateus C. Alves
Universidade Federal do Ceará
Departamento de Engenharia de Teleinformática
Fortaleza, Ceará, Brasil
vic.matteus@alu.ufc.br

Este trabalho apresenta uma análise exploratória de dados (EDA) do dataset "Heart Rate Prediction to Monitor Stress Level", com o objetivo principal de investigar os padrões que permitem a classificação dos níveis de estresse. O escopo da análise envolveu o pré-processamento dos dados, onde as classes de estresse (Baixo, Médio e Alto) foram definidas a partir da frequência cardíaca (HR). A metodologia seguiu uma abordagem estruturada, iniciando com análises univariadas (incondicionais e condicionais à classe) utilizando histogramas, box-plots e estatísticas descritivas. Subsequentemente, foram conduzidas análises bivariadas, incluindo matrizes de correlação e gráficos de dispersão, e uma análise de redução de dimensionalidade via Principal Component Analysis (PCA). Os resultados principais indicam que, embora o dataset apresente um bom potencial para distinguir as classes 'Alta' e 'Baixa', a classe 'Média' domina as dispersões, sugerindo um estado de transição fisiológica não estritamente definido. Variáveis como 'higuci fractal dimension of heartrate', 'sample entropy' e componentes de frequência (LF, VLF) mostraram-se relevantes para a separação das classes. Os insights obtidos demonstram o potencial do dataset para o desenvolvimento de modelos preditivos de estresse, ao mesmo tempo que destacam a complexidade da classe 'Média'

Keywords— Exploratory Data Analysis, Heart Rate Prediction and Stress Levels.

I. Introdução

O estresse é amplamente reconhecido como um dos principais fatores de risco para a saúde na sociedade moderna, estando associado a uma variedade de doenças crônicas, incluindo distúrbios cardiovasculares e problemas de saúde mental [Ref. 1]. A Organização Mundial da Saúde chegou a classificar o estresse como a "epidemia de saúde do século XXI", destacando a necessidade urgente de métodos eficazes para seu monitoramento e gerenciamento [Ref. 2]. Uma das formas não invasivas mais promissoras para quantificar o estresse fisiológico é através da análise da Variabilidade da Frequência Cardíaca (VFC), ou Heart Rate Variability. A VFC mede as flutuações no tempo entre batimentos cardíacos consecutivos e reflete a atividade do Sistema Nervoso Autônomo (SNA), que regula funções corporais involuntárias. O SNA é composto pelos ramos simpático (associado a respostas de luta ou fuga, ou estresse) e parassimpático (associado a respostas de descanso e digestão, ou relaxamento) [Ref. 3]. Níveis elevados de estresse estão tipicamente correlacionados a uma baixa variabilidade da frequência cardíaca, indicando um desequilíbrio autonômico onde a atividade simpática se sobrepõe à parassimpática [Ref. 4].

O dataset Heart Rate Prediction to Monitor Stress Level [Ref. 5] fornece um conjunto rico de métricas derivadas da VFC, permitindo uma análise profunda dessa relação. Além da frequência cardíaca (HR), o conjunto inclui métricas do domínio da frequência (como VLF, LF, HF e a razão LF/HF) e métricas não-lineares (como higuci e sampen), que analisam a complexidade do sinal cardíaco. A literatura sugere que tanto as métricas de frequência quanto as não-lineares são biomarcadores eficazes para a detecção de estresse [Ref. 6]. Contudo, antes que um modelo de machine learning robusto possa ser desenvolvido para classificar o estresse usando esses dados, uma Análise Exploratória de Dados (EDA) é fundamental. A EDA é necessária para compreender as características estatísticas do dataset, validar a qualidade dos dados, identificar as variáveis mais promissoras (features) e investigar a separabilidade das classes de estresse.

Este trabalho justifica-se pela necessidade de explorar e validar o potencial deste conjunto de dados específico para a classificação de estresse. As aplicações práticas de um modelo preditivo validado por esta análise são vastas, incluindo o monitoramento contínuo da saúde através de dispositivos wearables (como relógios e anéis inteligentes), o desenvolvimento de aplicativos de bem-estar que alertam sobre níveis de estresse cumulativo e a prevenção de condições como o burnout.

II. Métodos

A. Conjunto de Dados

O estudo utilizou o conjunto de dados público Heart Rate Prediction to Monitor Stress Level [Ref. 5], obtido da plataforma Kaggle. O dataset final, após o pré-processamento, é composto por 369.289 amostras (linhas) e 37 variáveis (colunas). As variáveis (features) abrangem um conjunto robusto de métricas da Variabilidade da Frequência Cardíaca (VFC), que podem ser agrupadas em três categorias principais:

- Domínio do Tempo: Métricas estatísticas que quantificam a variação nos intervalos RR (tempo entre batimentos cardíacos consecutivos), como MEAN_RR (Média dos intervalos RR), SDRR (Desvio padrão dos intervalos RR), RMSSD (Raiz quadrada da média das diferenças sucessivas ao quadrado) e pNN50 (Porcentagem de intervalos sucessivos com diferença > 50ms).

- Domínio da Frequência: Componentes de potência espectral que refletem a atividade do sistema nervoso autônomo, como VLF (Very Low Frequency), LF (Low Frequency) e HF (High Frequency), bem como suas proporções (LF_HF).

- Não-Lineares: Métricas que avaliam a complexidade e a regularidade da série temporal cardíaca, incluindo Sampen (Sample Entropy) e higuci (Dimensão Fractal de Higuchi). A variável

HR (Heart Rate) foi utilizada como base para a criação da variável-alvo.

B. *Pré-processamento e Definição das Classes*

O pré-processamento dos dados foi realizado em duas etapas principais. Primeiramente, os dados brutos, que estavam originalmente distribuídos em três arquivos separados (métricas de domínio do tempo, frequência e não-lineares), foram unificados em um único conjunto de dados por meio de scripts em Python. Variáveis de identificação, como *uuid* e *datasetId*, que não possuem valor preditivo, foram removidas da análise.

- Baixo: se HR <= 60 bpm
- Médio: se HR > 60 e < 100 bpm
- Alto: se HR >= 100 bpm

C. *Metodologia de Análise Exploratória (EDA)*

A metodologia de EDA foi dividida em três fases sequenciais para investigar a estrutura dos dados e a separabilidade das classes de estresse.

1. **Análise Univariada Incondicional**: Inicialmente, cada variável preditora foi analisada individualmente, sem considerar a classe. Para visualizar as distribuições, foram gerados histogramas e box-plots. Métricas de estatística descritiva (média, desvio padrão, mediana e skewness) foram calculadas para caracterizar o perfil de cada feature.

2. **Análise Univariada Condicional**: A análise anterior foi repetida, porém condicionada à classe de estresse (Baixo, Médio, Alto). Esta etapa visou observar como as distribuições de cada preditor se alteram entre os diferentes níveis de estresse, utilizando box-plots comparativos por classe.

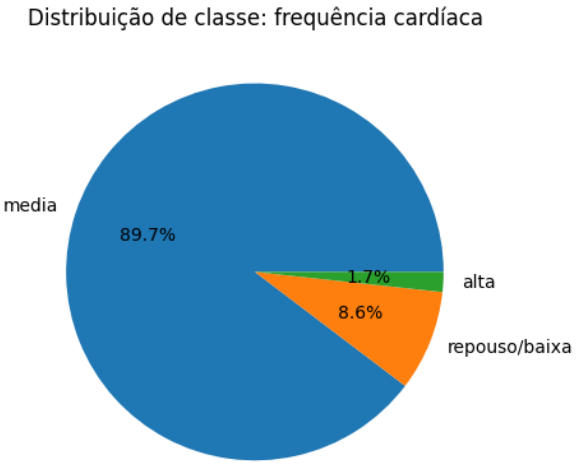
3. **Análise Bivariada e Multivariada**: A relação entre pares de variáveis foi investigada. Foi gerada uma matriz de correlação (utilizando o coeficiente de Pearson) e visualizada como um heatmap para identificar colinearidades e relações lineares com a HR. Gráficos de dispersão (scatter-plots) foram usados para examinar features promissoras. Por fim, aplicou-se a Análise de Componentes Principais (PCA) sobre os dados normalizados para reduzir a dimensionalidade e avaliar visualmente a separabilidade das três classes no espaço de features reduzido (plotando os dois primeiros componentes principais).

III. Resultados e Discussão

A. Análise de Classes e Análise Univariada Incondicional

A primeira etapa da análise foi a aplicação das regras de pré-processamento para definir a variável-alvo “classification”. Conforme definido na metodologia (HR <= 60 para 'repouso/baixa', HR >=100 para 'alta'), a distribuição das classes resultou em um conjunto de dados altamente desbalanceado. Dos 369.289 registros totais, a classe 'média' representa a esmagadora maioria, com 331.118 amostras, o que corresponde a 89,66% do total. As classes de extremo, 'repouso/baixa' e 'alta', são significativamente minoritárias, contando com 31.861 (8,63%) e 6.310 (1,71%) amostras, respectivamente. Esta distribuição é visualizada em detalhes na Figura 1, evidenciando o desbalanceamento dos dados quanto à distribuição de classes.

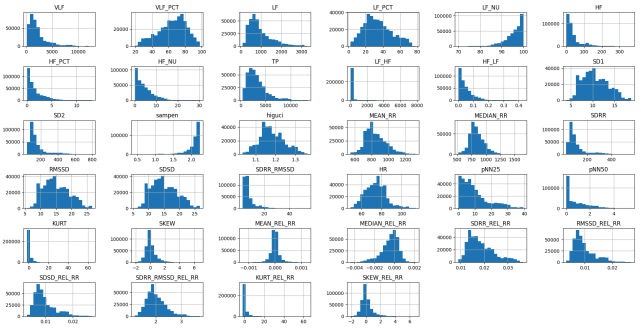
Figura 1: Distribuição da frequência cardíaca.



Fonte: Captura de tela do autor.

Esse desbalanceamento é um achado crítico, pois impacta diretamente a interpretação das análises e o eventual treinamento de modelos de classificação, que podem se tornar enviesados para a classe dominante. Em seguida, foi realizada a análise univariada incondicional de todas as features preditoras, cujas distribuições são apresentadas nos histogramas da Figura 2.

Figura 2: Distribuição dos histogramas para uma variável.



Fonte: Captura de tela do autor.

Observou-se que poucas variáveis seguem uma distribuição normal (Gaussiana). As features podem ser agrupadas por seu comportamento:

- **Distribuição Simétrica**: Um grupo de variáveis apresentou distribuições razoavelmente simétricas, aproximando-se da normalidade, como HR (Frequência Cardíaca), higuci, RMSSD e VLF_PCT.

- **Assimetria à Direita** (Positive Skew): Um número significativo de features apresentou forte assimetria à direita, com a maioria dos valores concentrada à esquerda (valores baixos). Este comportamento foi notável nas métricas de domínio de frequência, como VLF, LF, HF, HF_PCT e a razão LF_HF.

- **Assimetria à Esquerda** (Negative Skew): Um terceiro grupo exibiu assimetria à esquerda, com uma cauda de valores menores, observada em

métricas como MEAN_RR (Média dos intervalos RR) e HF_NU.

Esta análise inicial das distribuições gerais (incondicionais) sugere que as features possuem características estatísticas muito distintas, e a forte assimetria de muitas delas indica que transformações (como a logarítmica) podem ser necessárias em futuras etapas de modelagem.

B. *Análise Univariada Condicional*

A análise univariada foi então segmentada pela variável-alvo "classification", com o objetivo de verificar se as features preditoras apresentam distribuições distintas para os níveis de estresse 'Baixo', 'Médio' e 'Alto'. Os resultados desta análise, visualizados através de box-plots comparativos, disponíveis por meio da execução do programa desenvolvido, foram altamente reveladores.

Como esperado, a variável HR demonstrou uma separação perfeita, pois foi a feature utilizada para a criação das classes. Este resultado serve como uma validação do pré-processamento. De forma correlata, as métricas de domínio do tempo que são inversamente proporcionais à HR, especificamente MEAN_RR e MEDIAN_RR, também apresentaram uma separação ideal, com caixas totalmente distintas e medianas bem definidas para cada classe.

A descoberta mais significativa, no entanto, veio das métricas independentes, como as do domínio da frequência. Diversas variáveis demonstraram excelente potencial preditivo, apresentando pouca ou nenhuma sobreposição entre os intervalos interquartis (IQRs) das três classes:

- **Métricas de Frequência**: Variáveis como lf_nu, hf, hf_pct e hf_nu mostraram medianas claramente distintas. Por exemplo, em lf_nu e hf, o IQR (tamanho da caixa) da classe 'alta' era notavelmente grande, diminuindo drasticamente para a classe 'média' e tornando-se quase nulo para a 'baixa', indicando perfis de VFC muito diferentes.

- **Razões de Frequência**: As razões LF_HF e HF_LF também exibiram caixas sem sobreposição, sugerindo que o balanço simpatovagal é um forte diferenciador dos níveis de estresse definidos.

- **Métrica Relativa**: A feature SDRR_RMSSD_REL_RR destacou-se como um candidato particularmente forte, apresentando caixas pequenas, medianas centralizadas e separação total entre as classes, indicando alta previsibilidade.

Apesar da clara separação das medianas, uma observação crítica foi a alta incidência de outliers, especialmente nas classes 'média' e 'baixa'. Em features como lf_nu e lf_hf, a classe 'baixa' apresentou outliers superiores com um range muito superior ao seu próprio IQR. Adicionalmente, nem todas as features se mostraram robustas. A variável higuci, embora apresentasse medianas distintas, foi considerada menos confiável. Os whiskers (bigodes) da classe 'média' se estendiam por todas as classes, e seu IQR se sobrepunha parcialmente ao da classe 'alta'. Este achado reforça a observação do resumo: a classe 'média' parece representar um estado de transição fisiológico amplo e ruidoso, que se mistura com os estados de 'alto' e 'baixo' estresse.

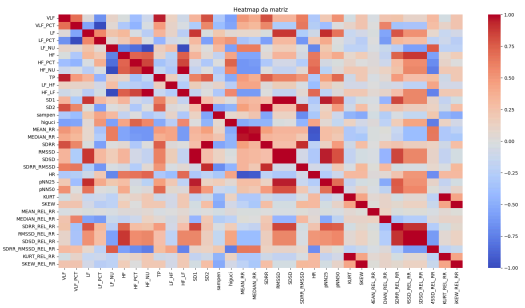
C. *Análise Bivariada e Multivariada*

Para finalizar a análise, foram investigadas as relações entre as variáveis (análise bivariada) e a separabilidade geral das classes em um espaço de *features* reduzido (análise multivariada).

1) *Análise de Correlação (Heatmap)*

A matriz de correlação, exposta na forma de heatmap na Figura 3, que calcula o coeficiente de Pearson entre todos os pares de *features*, revelou a existência de alta multicolinearidade no dataset.

Figura 3: Heatmap da matriz de correlação



Fonte: Captura de tela do autor.

Observam-se blocos de correlação muito forte (valores próximos de 1.0 ou -1.0):

- **Correlação de Validação**: Como esperado, as features de domínio do tempo, como MEAN_RR e MEDIAN_RR, apresentaram correlação positiva quase perfeita (próxima de 1.0) entre si, e correlação negativa quase perfeita (próxima de -1.0) com a HR. Isso é logicamente consistente, pois um aumento na frequência cardíaca (HR) significa uma diminuição no tempo entre batimentos (RR).

- **Multicolinearidade de Features**: Identificou-se alta correlação positiva entre features do mesmo domínio, como o bloco entre SDRR, RMSSD e SDSD.

- **Correlação com o Alvo**: De forma mais crítica para a predição, a HR (e, por extensão, a classe de estresse) mostrou correlações notáveis com features de outros domínios. Por exemplo, a HR apresentou correlação positiva com a razão LF_HF (associada à dominância simpática/estresse) e correlação negativa com HF (associada à atividade parassimpática/relaxamento).

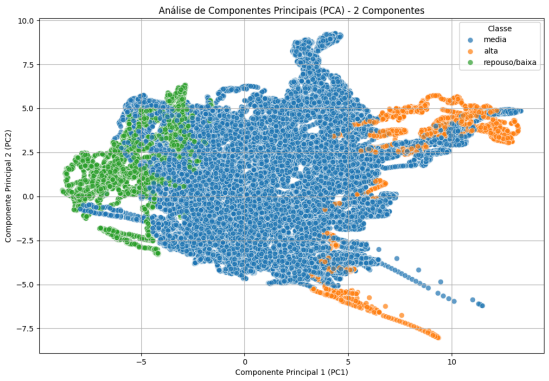
A forte multicolinearidade observada sugere que muitas *features* são redundantes. Isso justifica a aplicação de técnicas de redução de dimensionalidade, como o PCA, para criar componentes independentes e otimizar a modelagem futura.

2) *Análise de Componentes Principais (PCA)*

A análise de PCA foi aplicada aos dados normalizados para tentar visualizar a separabilidade das três classes de

estresse. O gráfico de dispersão dos dois primeiros componentes principais (PC1 e PC2) é apresentado na Figura 4.

Figura 4: Gráfico de dispersão do PCA



Fonte: Captura de tela do autor.

Este resultado visual é, talvez, o mais importante de toda a análise, pois confirma a hipótese central deste trabalho:

- **Separação dos Extremos:** As classes minoritárias, **'repouso/baixa' (verde)** e **'alta' (laranja)**, formam *clusters* (agrupamentos) visualmente distintos e bem separados nos extremos opostos do eixo do Componente Principal 1 (PC1). A classe 'baixa' está concentrada à esquerda (PC1 negativo), e a 'alta' está concentrada à direita (PC1 positivo).
- **A Classe "Média" como Transição:** A classe dominante **'média' (azul)** ocupa o vasto espaço central do gráfico. Crucialmente, ela não forma um *cluster* denso e isolado; em vez disso, ela atua como uma "ponte" difusa que se mescla com os limites dos *clusters* 'baixa' e 'alta'.

Esta visualização do PCA corrobora o que foi visto nos box-plots: enquanto os estados fisiológicos de 'alto' e 'baixo' estresse são distintos e caracterizáveis pelas *features* da VFC, o estado 'médio' é, de fato, uma zona de transição ampla, ruidosa e menos definida.

IV. Conclusão

Esta Análise Exploratória de Dados (EDA) investigou o dataset Heart Rate Prediction to Monitor Stress Level para avaliar seu potencial na classificação de níveis de estresse, definidos como 'Baixo', 'Médio' e 'Alto'. A análise revelou três padrões principais e um desafio crítico. Primeiramente, o principal problema identificado é o severo desbalanceamento de classes. A classe 'média' constitui aproximadamente 90% do conjunto de dados, o que representa um desafio técnico significativo para o desenvolvimento de modelos preditivos. Em segundo lugar, o principal padrão é que as classes de extremo ('baixa' e 'alta') são altamente distintas e separáveis. A análise univariada condicional demonstrou que diversas features da VFC — notavelmente métricas do domínio da frequência (como LF_NU, HF_PCT, LF_HF) e do domínio do tempo (MEAN_RR) exibem separação clara. Em terceiro lugar, o principal achado refere-se à natureza da classe 'média'. Tanto as análises condicionais quanto a visualização de redução de dimensionalidade (PCA) confirmaram que a classe 'média' não é um cluster coeso, mas sim um estado de transição fisiológico amplo e ruidoso, que atua como uma "ponte" difusa entre os estados de 'baixa' e 'alta'. Em conclusão, esta EDA confirma que o dataset possui features robustas.

Futuros trabalhos de modelagem devem focar em estratégias para lidar com o desbalanceamento extremo e reconhecer a dificuldade de definir uma fronteira nítida para o estado de estresse 'médio'.

V. Referências

1. A. Steptoe and M. Kivimäki, “Psychological stress and cardiovascular disease,” Nat. Rev. Cardiol., vol. 9, pp. 360–370, April 2012.

2. G. K. Saha, S. Sarkar, S. K. Roy, et al., “Stress: Prevalence and correlates among residents of a suburban area,” J. Family Med. Prim. Care, vol. 8, pp. 3844–3849, Dec. 2019.

3. P. Castiglioni, G. Parati, M. Di Rienzo, et al., “The role of heart rate variability (HRV) in different hypertensive syndromes,” Diagnostics, vol. 13, p. 744, Feb. 2023.

4. “Stress and heart rate variability: Relationship and management,” Medical News Today, Jan. 2024. [Online]. Available: https://www.medicalnewstoday.com/articles/stress-and-heart-rate-variability

5. V. Shanawad, “Heart Rate Prediction to Monitor Stress Level,” Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/vinayakshanawad/heart-rate-prediction-to-monitor-stress-level

6. “Heart Rate Variability: A Valuable Biomarker with A Major Impact...,” Hyperion Health, March 2024. [Online]. Available: https://www.hyperionhealth.ca/heart-rate-variability-a-valuable-biomarker-with-a-major-impact-on-physiological-and-psychological-health

7. I. Vitor Mateus Costa Alves, "Exploratory-Data-Analysis---HW1,"[Source code], 2025. [Online]. Available: VicMatteus/Exploratory-Data-Analysis---HW1: Análise exploratória de um conjunto de dados de monitoramento de ECG's para previsão de estresse.