Este documento inclui os textos desenvolvidos para o Artigo

Introdução |

O estresse pode ser definido como uma resposta adaptativa do organismo a pressões externas ou internas, envolvendo alterações neuroendócrinas e autonômicas que afetam o equilíbrio psicológico e fisiológico \cite{selye1950stress}. Embora seja uma reação natural, episódios prolongados ou intensos de estresse têm sido associados a condições de saúde adversas, incluindo o aumento da vulnerabilidade cardiovascular \cite{steptoe2012stress}. No Brasil, a relevância desse tema é reforçada por levantamentos que apontam que mais de metade da população adulta relata níveis elevados de estresse, frequentemente acompanhados de sintomas de ansiedade e depressão \cite{selye1950stress}.

No campo da ciência de dados, a exploração de sinais cardíacos em situações de estresse oferece uma oportunidade de identificar padrões relevantes para pesquisas em saúde e para o desenvolvimento de tecnologias de monitoramento. Medidas derivadas de sinais eletrocardiográficos (ECG), como a variabilidade da frequência cardíaca (HRV), têm se mostrado indicadores consistentes das alterações fisiológicas associadas ao estresse \cite{komorowski2016exploratory}.

Entre as bases públicas voltadas ao monitoramento fisiológico, destaca-se a \textit{Heart Rate Prediction to Monitor Stress Level}, disponibilizada por Shanawad \cite{Shanawad_HeartRate} no \textit{Kaggle}. O conjunto reúne medidas de variabilidade da frequência cardíaca (HRV) extraídas de sinais de ECG registrados em diferentes condições de estresse, permitindo investigar como alterações autonômicas se refletem em padrões estatísticos e dinâmicos dos batimentos cardíacos. Estudos recentes adotaram esse mesmo dataset para validar algoritmos de regressão e classificação. Izonin et al. \cite{izonin2023cascade} , por exemplo, propuseram um modelo cascade baseado em Support-Vector Regression que reduziu o erro de predição da frequência cardíaca em mais de 20× em relação a abordagens lineares. Já Silva et. al \cite{stacking2022} empregou stacking ensemble com seleção de características, alcançando F1-macro = 0,82 na classificação dos níveis de estresse.

Esses trabalhos demonstram o potencial do conjunto de dados para prever ou classificar níveis de estresse, mas não exploram as relações internas entre suas variáveis fisiológicas. Assim, compreender a estrutura estatística do dataset é essencial para avaliar sua consistência, redundância e relevância biológica antes da aplicação de modelos de aprendizado.

Diante disso, justifica-se a realização de uma análise exploratória de dados (Exploratory Data Analysis – EDA), etapa indispensável no processo analítico, que possibilita examinar distribuições, identificar valores atípicos, explorar relações entre variáveis e avaliar a consistência do conjunto antes da aplicação de métodos preditivos mais sofisticados. Nesse sentido, trata-se de um processo essencial para gerar hipóteses, guiar decisões metodológicas e fornecer uma visão estruturada sobre o comportamento dos dados \cite{data2016secondary}.

O presente trabalho tem como objetivo realizar um estudo exploratório do conjunto de dados anteriormente mencionado, \textit{Heart Rate Prediction to Monitor Stress Level}. Por meio de estatísticas descritivas, visualizações gráficas e análises multivariadas, busca-se compreender o comportamento das variáveis relacionadas à atividade cardíaca em diferentes condições de estresse. Ao fornecer uma visão inicial desses dados, este estudo contribui para a fundamentação de investigações futuras, seja no campo da modelagem preditiva, classificatória e, inclusive, no desenvolvimento de aplicações voltadas ao bem-estar e à saúde.

Metodologia |

Neste estudo, as análises foram realizadas em linguagem Python, com apoio de bibliotecas voltadas a manipulação e visualização de dados. O uso dessa linguagem se justifica pela sua versatilidade e popularidade em pesquisas acadêmicas, além de aplicações práticas. Todo o desenvolvimento foi versionado em repositório no GitHub \textcolor{red}{alguem referencia o git, pfv!!}, garantindo organização, rastreabilidade e reprodutibilidade dos experimentos conduzidos.

Conjunto de Dados

O conjunto de dados utilizado neste estudo é chamado \textit{Heart Rate Prediction to Monitor Stress Level}, disponibilizado publicamente na plataforma Kaggle. Ele reúne sinais de eletrocardiograma (ECG) processados em diferentes condições de estresse, a partir dos quais foram extraídos atributos em três domínios distintos: tempo, frequência e medidas não lineares.

%O conjunto de dados escolhido contém três arquivos CSV que abrangem diferentes tipos de características: características do domínio do tempo, características do domínio da frequência e características não lineares. Todos os arquivos CSV possuem a mesma quantidade de observações: 369.289, e não há nenhum valor ausente em todo o conjunto de dados.

No total, o dataset é composto por 369.289 observações, distribuídas em três arquivos para treinamento e três para teste, correspondentes aos diferentes grupos de atributos. Não há valores ausentes nos arquivos, o que favorece a consistência das análises exploratórias. Após a remoção dos identificadores técnicos (\texttt{uuid} e \texttt{datasetId}), o conjunto consolidado resulta em 33 preditores, que representam diferentes características estatísticas e dinâmicas dos sinais cardíacos extraídos do ECG. Dentre estas, uma variável categórica (\texttt{condition}), que descreve o estado do indivíduo no momento da medição, correspondendo a três classes: \textit{no stress} (200.082 instâncias), \textit{interruption} (105.150 instâncias) e \textit{time pressure} (64.057 instâncias); e \texttt{HR}, nossa variável de saída correspondente a frequência cardíaca registrada no momento da coleta.

%Como mencionado anteriormente, os campos "uuid" e "datasetId" são apenas identificadores para cada paciente e para o conjunto de dados, respectivamente. A variável "condition" é categórica, e "HR" é a variável de saída. Assim, o número total de preditores D é o número de colunas no conjunto

de dados combinado menos essas quatro, o que resulta em 33 preditores para este conjunto de dados.

%Existe apenas uma variável categórica em todo o conjunto de dados, "condition", portanto o número de classes L no conjunto é o número de categorias dessa variável, que são três: 'no stress', com 200.082 observações, 'interruption', com 105.150 observações, e 'time pressure', com 64.057 observações.

A diversidade de atributos, cobrindo domínios complementares da atividade cardíaca, fornece um conjunto rico para investigação inicial da análise exploratória de dados (EDA). Esse cenário é particularmente adequado para examinar a relação entre estresse e variabilidade da frequência cardíaca, além de permitir a identificação de padrões, outliers e potenciais fatores discriminativos entre condições.

Referenciar tabela!!
"o conjunto consolidado resulta em 33 variáveis, que representam diferentes características estatísticas e dinâmicas dos sinais cardíacos extraídos do ECG, sendo as principais estão apresentadas na Tabela \ref{tab:descricao_data}. "

Análise Multivariada

Para complementar a análise univariada e bivariada dos preditores, foi conduzida também uma análise multivariada do conjunto de dados, utilizando a técnica de Análise de Componentes Principais (Principal Components Analysis – PCA). A PCA é um método estatístico que permite projetar dados de alta dimensão em um espaço de menor dimensionalidade, preservando a maior parte da variância original \cite{mackiewicz1993principal}. Tal procedimento possibilita identificar padrões globais, correlações entre variáveis e potenciais agrupamentos de observações, ao mesmo tempo em que facilita a visualização dos dados.

Neste trabalho, a PCA foi implementada manualmente, sem o uso direto de funções pré-existentes, de modo a reforçar a compreensão do método. O processo consistiu nas seguintes cinco etapas: (i) pré-processamento dos dados, com centralização e padronização das variáveis; (ii) cálculo da matriz de covariância entre os preditores; (iii) obtenção dos autovalores e autovetores dessa matriz; (iv) seleção dos autovetores associados aos maiores autovalores, correspondentes às componentes principais de maior variância explicada; e (v) projeção dos dados no novo espaço de características, resultando em uma representação de dimensionalidade reduzida.

Para avaliar a importância relativa de cada componente principal, empregou-se inicialmente o \textit{scree plot}, que representa a variância explicada por cada componente em ordem decrescente. A variância explicada, por sua vez, indica a proporção da variabilidade total dos dados que é capturada por cada componente, permitindo identificar o número de dimensões que mais contribuem para representar o conjunto sem perda substancial de informação.

Na sequência, foi também estudado o \textit{scatter plot} bidimensional com as observações projetadas nas duas primeiras componentes principais. Essa visualização foi analisada sob duas perspectivas complementares: (i) uma categórica, baseada na variável \texttt{condition}, que

permite examinar a distribuição das diferentes classes de estresse no espaço reduzido; e (ii) uma contínua, fundamentada na variável \texttt{HR} (frequência cardíaca), em que as observações são graduadas conforme seus valores numéricos. Essa segunda abordagem não substitui a categórica, mas a complementa ao revelar variações graduais e relações contínuas entre a frequência cardíaca e as direções de maior variabilidade nos dados, oferecendo uma dimensão fisiológica adicional à análise.

Assim, a combinação entre o \textit{scree plot} e as projeções bidimensionais obtidas via \textit{scatter plot} fornece uma visão abrangente da estrutura multivariada dos dados, integrando tanto a distribuição global das variáveis quanto os padrões de organização observados no espaço das componentes principais..

Resultados | Análise Multivariada

A análise dos \textit{scree plots} permitiu avaliar a proporção de variância explicada por cada componente principal (Figuras~\ref{fig:scree_condition} e~\ref{fig:scree_hr}). Observou-se que as duas primeiras componentes capturam aproximadamente metade da variabilidade total do conjunto: 27,93\% e 25,13\% para o caso categórico (\texttt{condition}), e 26,80\% e 25,68\% para o caso contínuo (\texttt{HR}), totalizando 53,06\% e 52,49\%, respectivamente. O ponto de inflexão observado nas curvas acumuladas sugere que aproximadamente quatro componentes seriam suficientes para reter cerca de 80\% da variância total, o que indica uma estrutura de dados relativamente concentrada nas primeiras dimensões. Considerando, porém, o objetivo exploratório e de visualização, optou-se por representar os dados nas duas primeiras componentes principais, que concentram a maior parcela da variância explicada e permitem observar de forma mais intuitiva os padrões globais do conjunto.

Quando as observações foram coloridas pelas condições de estresse (Figura~\ref{fig:pca_condition}), constatou-se que as classes não estão bem separadas. Todas apresentam ampla sobreposição, sem evidência de fronteiras lineares ou agrupamentos claramente definidos. Observa-se, ainda, que a classe \textit{no stress} ocupa uma região mais extensa do espaço projetado, efeito esperado pelo maior número de instâncias, mas que pode gerar a impressão de maior dispersão em relação às demais. Assim, o resultado sugere que as variáveis disponíveis disponíveis capturam de forma limitada as diferenças entre os estados de estresse, o que dificulta a construção de modelos classificatórios lineares.

Em contraste, quando HR foi representada como gradiente de cor (Figura~\ref{fig:pca_hr}), surgiu um padrão contínuo e bem definido ao longo da CP1. Valores baixos concentraram-se em uma extremidade, enquanto valores altos se organizaram na oposta. Essa estrutura evidencia uma forte correlação entre HR e a primeiro componente, corroborada pela análise das cargas fatoriais: variáveis de ritmo médio e atividade espectral de alta frequência (MEAN-RR, HF, HF-PCT) contribuíram fortemente para a CP1, enquanto métricas de variabilidade de curto prazo (RMSSD, SDSD, SD1) influenciaram principalmente a CP2.

Assim, a PCA evidencia uma dualidade relevante. Enquanto os níveis de estresse categóricos não se separam de maneira nítida, mostrando ampla sobreposição e ausência de fronteiras lineares, a

frequência cardíaca apresenta um padrão contínuo e bem definido ao longo da primeira componente. Esses resultados sugerem que os preditores são mais adequados para tarefas de regressão da frequência cardíaca do que para a classificação direta dos níveis de estresse.

Do ponto de vista fisiológico, os resultados indicam que a resposta autonômica ao estresse não se manifesta de forma discreta, mas sim como um contínuo de variações nos parâmetros cardíacos. O aumento gradual da frequência cardíaca acompanha a intensificação do estado de estresse, refletindo uma organização linear ao longo da primeira componente. Assim, a análise multivariada não apenas descreve a estrutura estatística do conjunto, mas também fornece insights sobre a natureza integrada e progressiva das respostas fisiológicas, oferecendo subsídios valiosos para estratégias de monitoramento e quantificação do estresse a partir de sinais cardíacos.