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}.

Para alcançar tal compreensão inicial, a análise exploratória de dados (Exploratory Data Analysis – EDA) desempenha papel central. Considerada uma etapa indispensável no processo analítico, a EDA 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 \textit{Heart Rate Prediction to Monitor Stress Level}, disponível na plataforma Kaggle. 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, classificatórias 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.

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, neste caso realizada em duas dimensões.

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 dois autovetores correspondentes aos maiores autovalores, ou seja, as duas primeiras componentes principais; e (v) projeção dos dados no novo espaço de características, resultando em uma representação de dimensionalidade reduzida adequada para visualização bidimensional.

Para interpretação exploratória dos resultados, foram consideradas duas perspectivas de visualização. Na primeira, as observações foram representadas em função da variável categórica \texttt{condition}, que indica o estado de estresse dos indivíduos (Figura~\ref{fig:pca_condition}). Essa abordagem permite examinar como as diferentes classes se distribuem no espaço das duas primeiras componentes principais. Na segunda, a variável numérica \texttt{HR} (frequência cardíaca) foi utilizada como referência visual, sendo representada em gradiente de cor (Figura~\ref{fig:pca_hr}). Essa abordagem busca explorar a relação contínua entre a frequência cardíaca e as direções de maior variabilidade nos dados.

Dessa forma, as duas visualizações complementam-se: a primeira fornece uma análise baseada em categorias de estresse, enquanto a segunda adota uma perspectiva quantitativa associada a uma medida fisiológica contínua.

\begin{figure}[h]

\centering

\includegraphics[width=0.48\textwidth]{Figuras/PCA Visualizacao por Nivel de Estresse.png}

\caption{Projeção das observações nas duas primeiras componentes principais, coloridas segundo a variável categórica \texttt{condition} (nível de estresse).}

\label{fig:pca_condition}

\end{figure}

\begin{figure}[h]

\centering

\includegraphics[width=0.48\textwidth]{Figuras/PCA Visualizacao por Frequencia Cardiaca.png}

\caption{Projeção das observações nas duas primeiras componentes principais, representando a variável contínua \texttt{HR} (frequência cardíaca) em gradiente de cor.}

\label{fig:pca_hr}

\end{figure}

Resultados

A projeção dos dados nas duas primeiras componentes principais (CP1 e CP2), obtidas a partir da Análise de Componentes Principais (PCA), revelou padrões distintos conforme a variável de referência adotada. Para as classes de estresse (\texttt{condition}), a CP1 explicou 27,93\% da variância e a CP2, 25,13\%, totalizando 53,06\%. Para a frequência cardíaca (HR), os valores foram semelhantes, 26,80\% e 25,68\%, acumulando 52,49\%. Esses percentuais indicam que duas dimensões capturam mais da metade da variabilidade do conjunto, embora parte relevante ainda permaneça distribuída em componentes de ordem superior, que devem ser úteis em aplicações futuras.

Quando as observações foram coloridas pelas condições de estresse, constatou-se que as classes não estão bem separadas. Todas apresentam ampla sobreposição, sem evidência de fronteiras lineares ou agrupamentos definidos. Em particular, a classe \textit{no stress} mostrou-se altamente misturada tanto com \textit{interruption} quanto com \textit{time pressure}, configurando-se como a mais difícil de distinguir. Esse resultado reforça que os preditores disponíveis não capturam de forma clara as diferenças entre níveis de estresse, o que impõe limitações relevantes para classificadores supervisionados, sobretudo os lineares.

Em contraste, quando HR foi representada como gradiente de cor, emergiu 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 revela um contraste importante. Enquanto as classes de estresse apresentam fraca separação, ausência de limites lineares e alta sobreposição, especialmente envolvendo a condição \textit{no stress}, a HR é representada de forma consistente e contínua. Esses resultados indicam que os preditores são mais adequados para tarefas de regressão da frequência cardíaca do que para a classificação direta dos níveis de estresse.