Lucas da Fonseca Costa rm560575

O dataset de dados numéricos escolhido foi o **Heart Failure Prediction Dataset** que consiste em um agrupamento de 5 datasets que foram combinados com base em 11 características comuns. Os cinco conjuntos de dados usados para sua curadoria são:

- Cleveland: 303 registros
- Hungarian: 294 registros
- Switzerland: 123 registros
- Long Beach VA: 200 registros
- Stalog (Heart) Data Set: 270 registros

No total, o dataset reúne **1.190 observações**. Durante o processo de curadoria, foram identificadas **272 duplicações**, resultando em **918 registros únicos**, que constituem a base final para análise e modelagem preditiva.

Para as variáveis que acredito são as mais relevantes são:

ChestPainType (Tipo de dor torácica) - Sintoma central na avaliação cardiológica, especialmente angina típica. É um dos indicadores mais fortes para investigação de doença arterial coronariana.

ExerciseAngina (angina induzida por esforço) - Mostra limitação funcional e isquemia desencadeada pelo esforço. Relevante para prognóstico e estratificação de risco.

MaxHR (frequência cardíaca máxima atingida) - Importante em teste ergométrico. Baixa resposta cronotrópica pode indicar doença cardíaca.

RestingBP (pressão arterial de repouso) - Hipertensão é um dos principais fatores de risco cardiovascular.

Cholesterol e **FastingBS** (perfil metabólico) - Hipertensão, diabetes e dislipidemia são fatores de risco fundamentais para aterosclerose.

Para a análise de dados visuais, utilizei o **dataset CardiacUDC** , que contém diversos vídeos de ecocardiogramas no formato .nii. A fim de preparar esses dados para processamento, converti os vídeos em slices (fatias) individuais no formato .png, utilizando um script desenvolvido e executado no Google Colab.

Detecção de padrões como movimento das paredes cardíacas e variações estruturais

Identificação de bordas e contornos como delimitação das câmaras cardíacas e válvulas

Reconhecimento de anomalias como hipertrofia, dilatação e alterações de contratilidade

Estas análises são de grande relevância para projetos de IA aplicada na saúde, pois facilitam a criação de sistemas capazes de auxiliar no diagnóstico precoce, especialmente em doenças cardiovasculares, que estão entre as principais causas de morbidade e mortalidade no mundo.

Para o processamento textual foi escolhido o texto **Estatística Cardiovascular – Brasil 2023** que tem como objetivo fornecer uma compilação anual dos dados e das pesquisas sobre a epidemiologia das DCV no Brasil. Este documento congrega as estatísticas oficiais do Ministério da Saúde do Brasil e outras entidades governamentais ao lado de dados do projeto GBD, coordenado pelo IHME da Universidade de Washington

Por se tratar de uma fonte que consolida dados atualizados de forma sistemática e anual, esse documento representa uma excelente oportunidade para ser explorado por algoritmos de NLP.

E o segundo texto escolhido foi o **Distribution and spatial autocorrelation of the hospitalizations for cardiovascular diseases in adults in Brazil** que é um estudo ecológico das taxas de internação por doenças cardiovasculares em adultos no período de 2005 a 2016. A dependência espacial foi analisada pelo coeficiente de autocorrelação de Moran Global e Local.

Esse tipo de texto pode ser explorado por algoritmos de NLP através da análise de sentimentos, permitindo identificar se há tendências de aumento ou redução das doenças cardíacas ao longo do tempo. Além disso, é possível aplicar técnicas de extração de informações para identificar as regiões com maiores índices de hospitalização e os principais fatores de risco associado