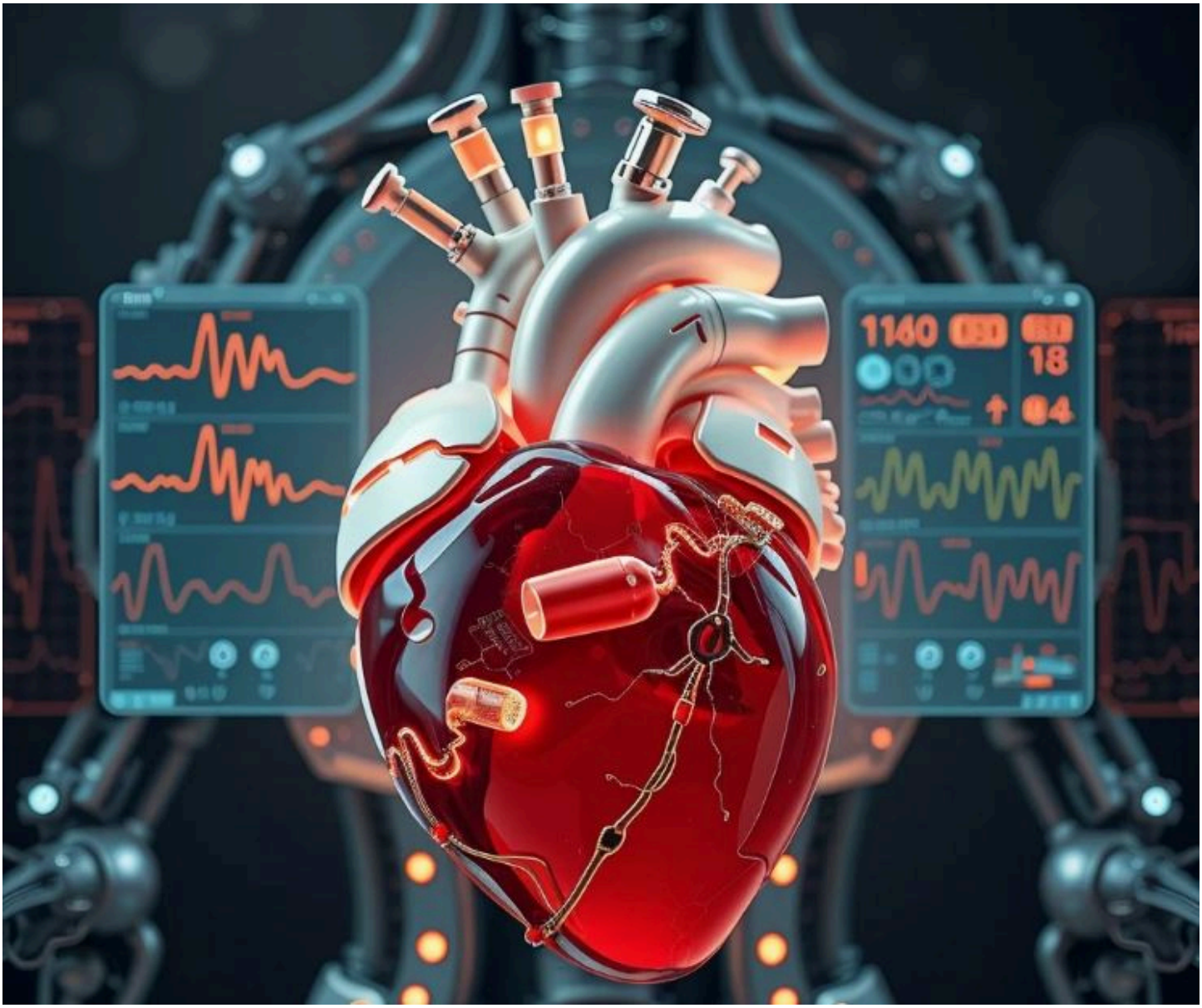
PROJETO BATIMENTO DE DADOS:

MAPEANDO O CORAÇÃO MODERNO



SOBRE O PROJETO

As doenças cardiovasculares (DCV) representam a principal causa de mortalidade e morbidade em escala global, um desafio de saúde pública que transcende fronteiras e sistemas de saúde.

De acordo com a Organização Mundial da Saúde (OMS), as DCV são responsáveis por aproximadamente 19,8 milhões de mortes por ano no mundo, o que as coloca no topo da lista das causas de óbito. No Brasil, o cenário não é diferente: em 2022, as doenças cardiovasculares foram responsáveis pela perda de quase 400.000 vidas.

Essa alta incidência e mortalidade estão diretamente ligadas a uma combinação de fatores de risco. Eles são classicamente divididos em duas categorias:

- Fatores de Risco Modificáveis:

São aqueles que podem ser prevenidos ou controlados por meio de intervenções no estilo de vida e tratamento médico. Incluem hipertensão, diabetes, níveis elevados de colesterol, tabagismo, sedentarismo, obesidade e má alimentação. A detecção precoce e o manejo desses fatores são cruciais para a prevenção e redução da carga das DCV.

- Fatores de Risco Não Modificáveis:

São características inerentes ao indivíduo, como idade, gênero, histórico familiar e predisposição genética. Embora não possam ser alterados, o conhecimento desses fatores é vital para a identificação de grupos de alto risco, permitindo uma vigilância e intervenção clínica mais focada.

Nesse contexto, o projeto CardioIA surge como uma resposta inovadora a esse desafio. Ele simula um ecossistema de cardiologia inteligente, unindo o poder da Inteligência Artificial (IA) com a análise de dados clínicos, visuais e textuais. Nosso objetivo é não apenas entender a prevalência e a mortalidade das DCV, mas também construir ferramentas que possam auxiliar na triagem,

diagnóstico, monitoramento e previsão de riscos, contribuindo para um futuro mais saudável.

A jornada do projeto, iniciada nesta Fase 1 - Batimentos de Dados, tem como objetivo a coleta e preparação das bases de dados. Esta etapa é fundamental, pois é o alicerce sobre o qual todos os módulos futuros de Machine Learning, Visão Computacional e Processamento de Linguagem Natural serão desenvolvidos.

GOVERNANÇA E SEGURANÇA DOS DADOS

Este projeto foi construído com base nas seguintes práticas, garantindo total conformidade com a Lei Geral de Proteção de Dados (LGPD):

- Anonimato dos Dados: O dataset utilizado é público, proveniente do Kaggle, e já passou por um processo de anonimização.
- Não Coleta de Dados Pessoais: Não houve manipulação de dados sensíveis. Cada paciente é representado por um código de identificação único, garantindo a privacidade total.
- Integridade e Confidencialidade: Esta abordagem nos permite realizar a análise de dados cardiológicos de forma segura e ética, mantendo a integridade das informações e a confidencialidade dos pacientes.

DATASET NUMÉRICO

O dataset original é composto por por 12 colunas e dividido em 3 categorias de variáveis:

👤 **Variáveis Demográficas:**

- `id`: Identificador único do paciente.
- `age`: Idade do paciente em dias.
- `gender`: Sexo do paciente (1 = feminino, 2 = masculino).

- height: Altura do paciente em centímetros.
- weight: Peso do paciente em quilogramas.

🩺 **Variáveis de Exame:**

- ap_hi: Pressão arterial sistólica em mmHg.
- ap_lo: Pressão arterial diastólica em mmHg.
- cholesterol: Nível de colesterol (1: normal, 2: acima do normal, 3: muito acima do normal).
- gluc: Nível de glicose (1: normal, 2: acima do normal, 3: muito acima do normal).
- smoke: Se o paciente fuma (0: não, 1: sim).
- alco: Se o paciente consome álcool (0: não, 1: sim).
- active: Nível de atividade física (0: não ativo, 1: ativo).

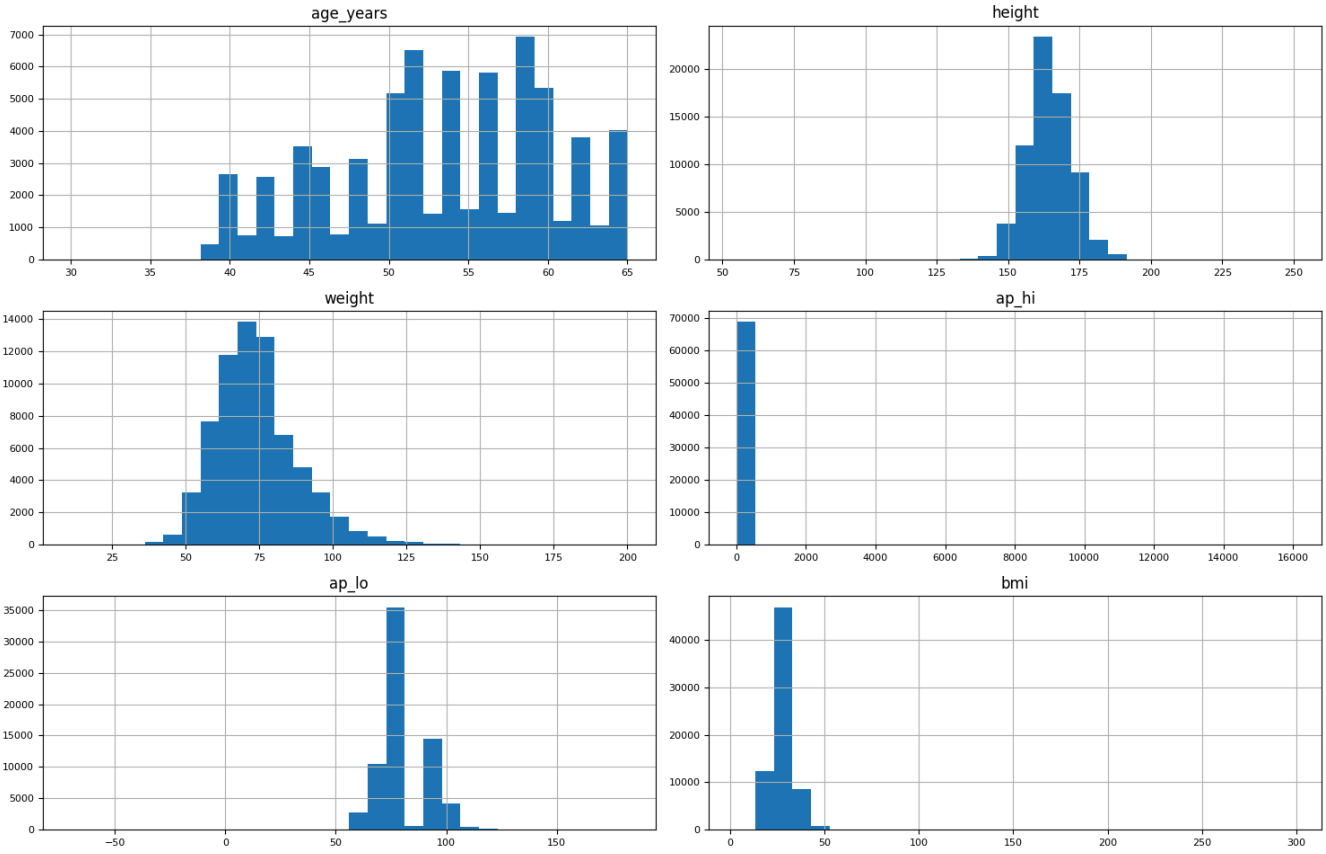
🎯 **Variável-alvo**

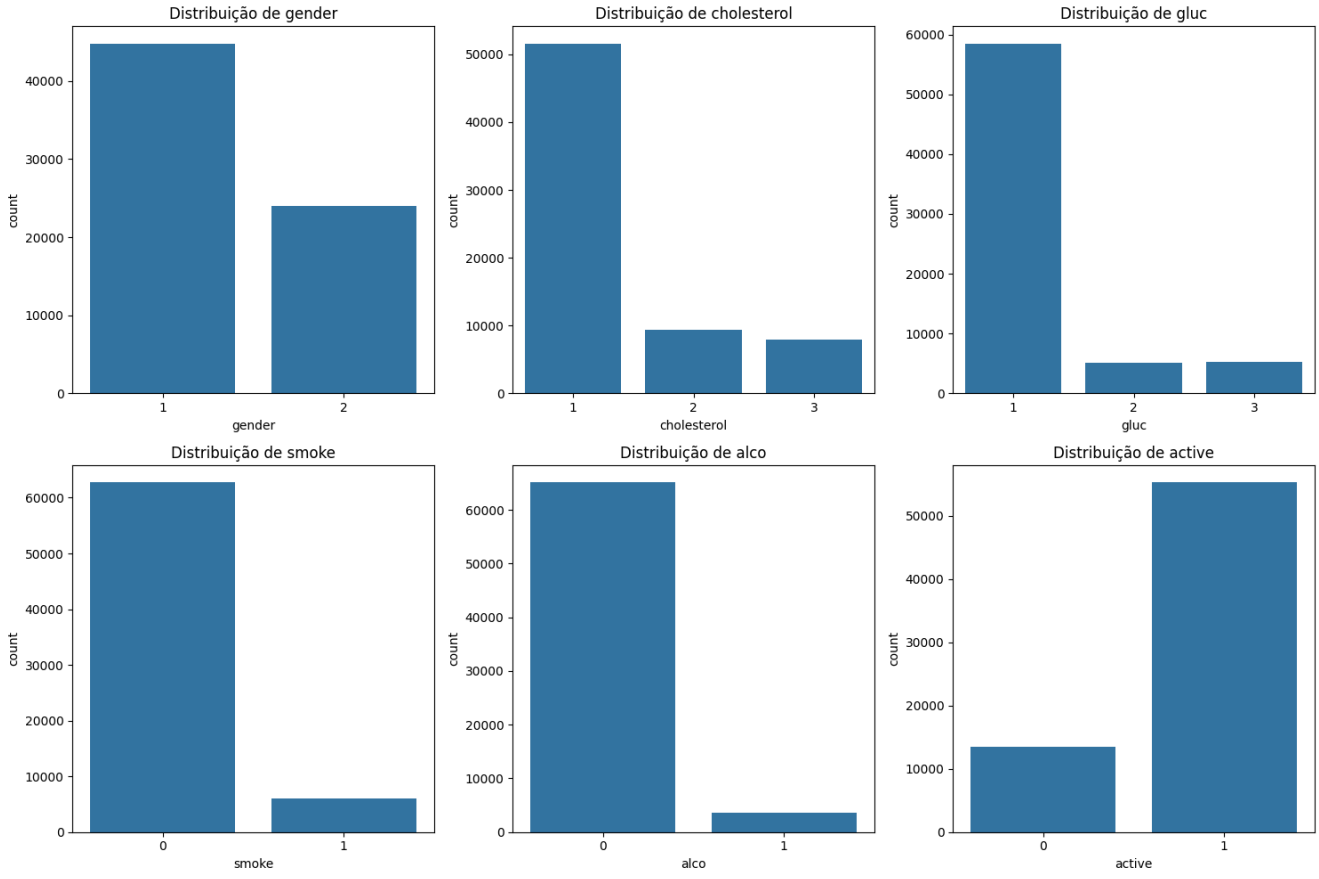
- cardio: indica a presença ou ausência de doença cardiovascular (0: ausente, 1: presente).

JUSTIFICATIVA PARA A ESCOLHA DO DATASET

A escolha do dataset ***cardio_train.csv*** para o projeto "Batimentos de Dados" é justificada pela sua relevância direta e pela riqueza de informações que oferece para a análise e modelagem de doenças cardiovasculares.

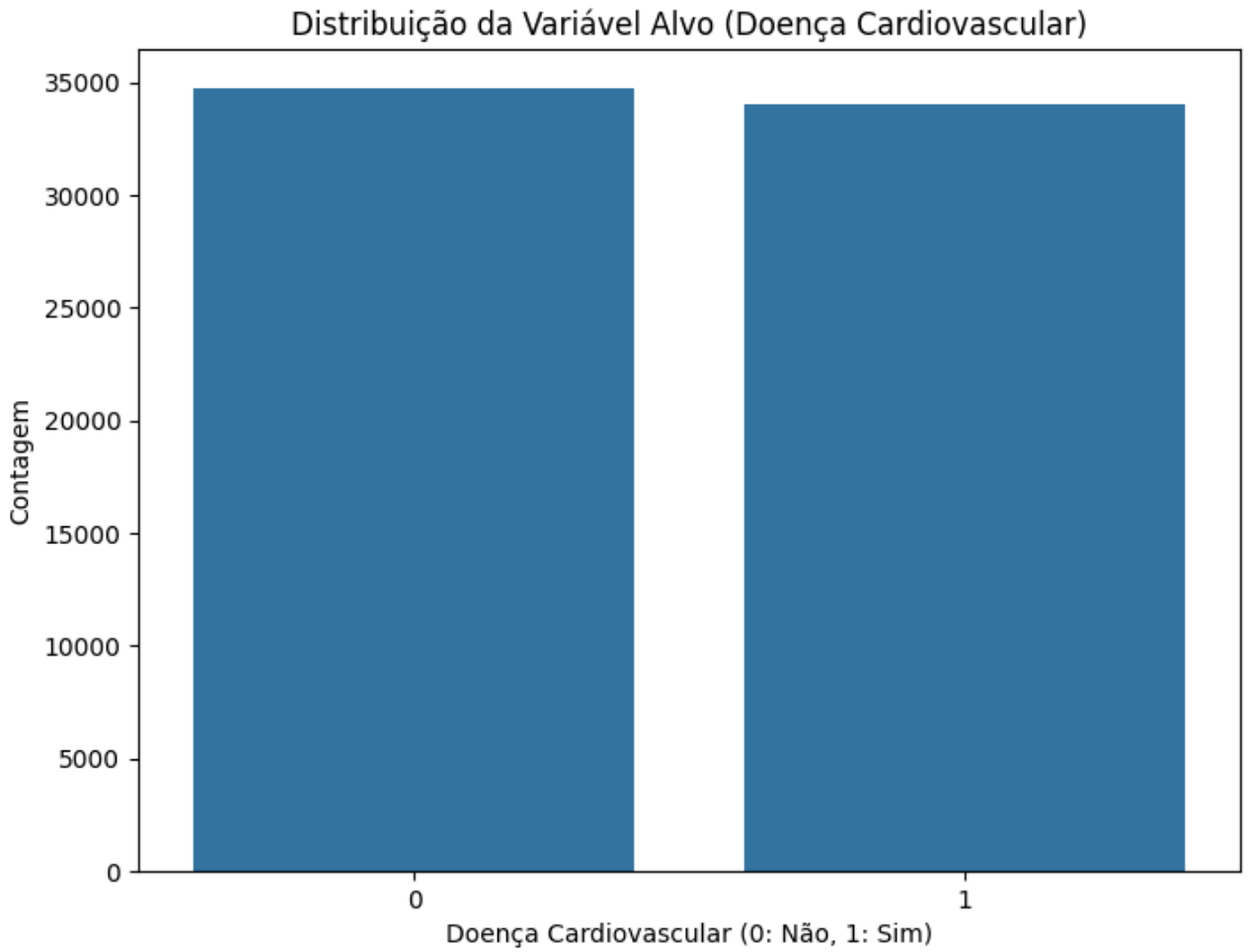
A combinação de dados demográficos, clínicos e de estilo de vida e a inclusão de variáveis como idade, gênero, peso e altura fornecem um conjunto de dados robusto e adequado para treinamento em ML/DL, possibilitando a identificação de padrões e precisão do risco de adoecimento cardiovascular, fornecendo insights valiosos.


age_years
height
weight
ap_hi
ap_lo
bmi


Distribuição de gender
gender
count
Distribuição de cholesterol
cholesterol
count
Distribuição de gluc
gluc
count
Distribuição de smoke
smoke
count
Distribuição de alco
alco
count
Distribuição de active
active
count

Suas métricas clínicas essenciais, como pressão arterial (ap_hi, ap_lo), colesterol e glicose, e de dados sobre hábitos de vida (fumo, álcool e atividade física) permite que o projeto vá além de uma análise puramente clínica e explore o impacto de fatores modificáveis na saúde do coração.

A variável de destino (cardio) é binária, indicando a presença ou ausência de doença cardiovascular, o que torna o dataset perfeitamente adequado para a construção de um modelo de classificação.



DATASETS TEXTUAIS

Análise de Textos Médicos com NLP

Este repositório demonstra como **técnicas de Processamento de Linguagem Natural (NLP)** em artigos científicos sobre **doenças cardiovasculares** e fatores de risco associados podem ser explorados. O objetivo é extrair insights

relevantes que apoiem projetos de **Inteligência Artificial em Saúde**, contribuindo para prevenção, diagnóstico e políticas públicas.

Artigos Utilizados

1. Inflamação sistêmica causada pela periodontite crônica em pacientes vítimas de ataque cardíaco isquêmico agudo
   - Estuda a associação entre periodontite, inflamação sistêmica e risco de infarto.
   - Contém dados clínicos: LDL, HDL, triglicerídeos, glicemia, e agentes biológicos como bactérias periodontais.
2. Associação entre saúde cardiovascular e depressão autorreferida: Pesquisa Nacional de Saúde 2019
   - Explora a relação entre saúde mental (depressão) e indicadores de saúde cardiovascular: IMC, tabagismo, hipertensão e diabetes.
   - Base populacional: 57.898 adultos brasileiros.

Aplicações de NLP

1. Extração de Entidades Médicas (NER)

- Identificação automática de **biomarcadores**: LDL, HDL, glicemia.
- Reconhecimento de **agentes infecciosos**: *Porphyromonas gingivalis*, *Prevotella intermedia*.
- Extração de **condições clínicas**: hipertensão, diabetes, depressão.
- Mapeamento para terminologias padronizadas como **SNOMED-CT** e **UMLS**.

Benefício: Permite estruturar informações clínicas de textos livres, facilitando integração com **bancos de dados de saúde** e **sistemas clínicos**.

2. Classificação de Tópicos

- **Artigo 1:** *doenças periodontais*, *inflamação sistêmica*, *cardiopatias isquêmicas*.
- **Artigo 2:** *fatores biológicos*, *hábitos comportamentais*, *saúde mental*.

Benefício: Organiza automaticamente a literatura científica, apoiando **pesquisadores** na identificação de conexões entre condições de saúde.

3. Mineração de Relações Causais

- Identificação de padrões como:
  - *"Periodontite crônica → inflamação → risco de infarto"* (Artigo 1).
  - *"Depressão → maior prevalência de doenças cardiovasculares"* (Artigo 2).

Benefício: Fundamenta **modelos explicáveis de risco clínico**, essenciais para **IA interpretável em saúde**.

4. Análise de Sentimentos e Autorrelatos

- Processamento de depoimentos de pacientes para detectar indícios de **tristeza, ansiedade ou risco de depressão**.
- Classificação automática de relatos em **neutros**, **positivos** ou **depressivos**.

Benefício: Suporta **triagem populacional** e programas de **atenção primária à saúde mental**.

5. Integração com Modelos Preditivos

- Dados extraídos dos textos podem alimentar **modelos de Machine Learning** para prever risco cardiovascular ou depressão.
- Exemplo: pacientes com periodontite crônica grave + LDL elevado → maior probabilidade de infarto.

Benefício: Viabiliza **sistemas de apoio à decisão clínica** e recomendações preventivas.

Conclusão

A aplicação de NLP aos artigos selecionados evidencia como a **Inteligência Artificial pode transformar textos científicos em bases de conhecimento estruturadas**, permitindo:

- Diagnóstico precoce de doenças cardiovasculares e mentais.
- Apoio à **formulação de políticas públicas** de saúde.
- Integração da **saúde física e mental** do paciente.
- Avanço científico com insights de alto valor clínico.

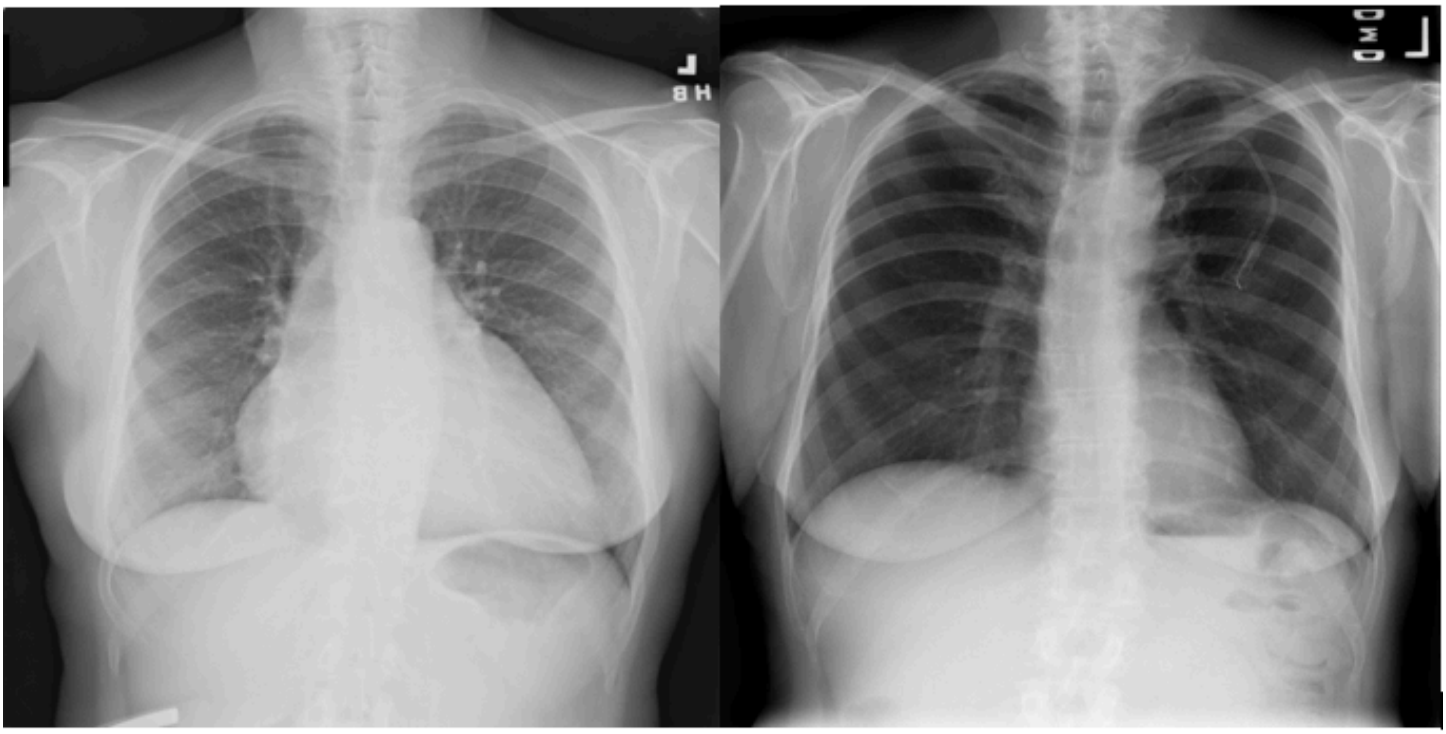
DADOS VISUAIS

Para esse projeto o dataset selecionado está relacionado a radiografias de tórax por diversas razões:

- A radiografia de tórax é uma ferramenta amplamente disponivel e de baixo custo para triagem e estratificação da doença cardíaca. Esses exames possibilitam uma visualização direta da silhueta cardíaca em relação à cavidade torácica. Isso permite a implementação de uma tarefa de VC clara e clinicamente relevante: a detecção de cardiomegalia (aumento do

coração) através do cálculo da Relação Cardiotorácica (RCT). A RCT é uma métrica estabelecida que os radiologistas usam, tornando sua automação um exemplo exemplar de como os algoritmos de VC podem replicar e padronizar a análise diagnóstica. A tarefa envolve a segmentação de estruturas anatômicas (coração e tórax), extração de características (diâmetros máximos) e classificação baseada em regras (RCT > 0.5), que se alinham perfeitamente com os princípios básicos da visão computacional.

- Grandes conjuntos de dados públicos e bem documentados, como o NIH ChestX-ray14 e o CheXpert de Stanford, estão prontamente disponíveis para uso em pesquisa acadêmica. Esses repositórios contêm dezenas a centenas de milhares de imagens, muitas já em formatos de imagem padrão como PNG ou JPG, eliminando a necessidade de conversão complexa de formatos.



Para extrair as imagens do dataset original foi criado um código em python que selecionou 200 imagens (100 com a feature ‘cardiomegalia’ e 100 com a feature ‘sem cardiomegalia’).

Esse código pode ser visualizado na integra através do link abaixo:

https://www.kaggle.com/code/iolandahfmanzali/fiap-f1-25?scriptVersionId=257394881