Análise Exploratória de Dados aplicada a Diagnósticos de Doenças Cardíacas, Diabetes e Câncer.

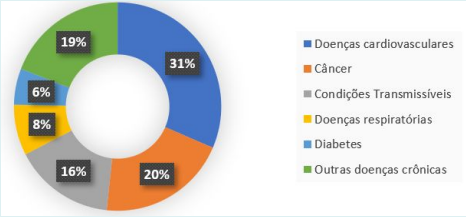
Gustavo Z. do Carmo, Martin Ropke, Raphael M. Calciolari.

Destaques

- Dados iniciais: Três bases de dados independentes, com variáveis sem nenhuma relação aparente.
- Proposta: Fazer o tratamento de dados dos *datasets* e encontrar possíveis relações entre as suas variáveis.
- Identificar possíveis fatores que indiquem uma tendência ao diagnóstico positivo das doenças.

Motivação

Dentre as doenças que afetam o Brasil, as doenças cardiovasculares, a diabetes e o câncer representam, na distribuição das causas de óbito, cerca de 51% dos casos.



Dados

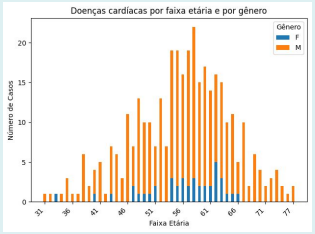
Todos os *datasets* foram obtidos do site IEEEDataPort

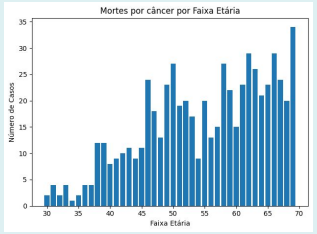
Desafios

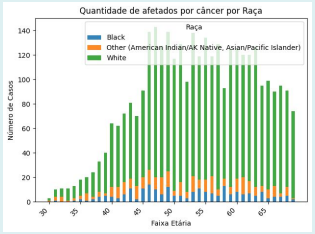
Um dos maiores desafios encontrados no desenvolvimento da EDA foi encontrar uma relação entre as variáveis que compunham os *datasets,* pois os dados selecionados apresentavam muitas variáveis. Além disso, os dados não estavam propriamente tratados, sendo necessária a remoção de células com valores nulos ou duplicados.

Fluxograma da Análise

Resultados







A faixa etária mais afetada pelas doenças é, em média, 53 anos, e as pessoas de 69 anos são as que mais correm risco de morte por câncer.

Correlações

A massa corporal e o açúcar médio no sangue são variáveis importantes para a classificação da diabetes, assim como a idade. Nos casos de doença de coração é notável a relevância da inclinação do ST, que representa o começo da repolarização ventricular, além da presença de angina(dor no coração) após exercícios físicos. No caso do câncer, o fato de estar localizado em uma parte do corpo ou espalhado é um grande fator para definir as chances de vida.