

# Análise de Dados sobre Diabetes : Storytelling

## Introdução

Esse projeto nasceu da curiosidade em entender como diferentes variáveis clínicas se relacionam com o diabetes. Para isso, analisei um conjunto de dados com informações biomédicas de mil pacientes, usando técnicas de Análise Exploratória de Dados (EDA) para investigar padrões, levantar hipóteses e refletir sobre os possíveis fatores de risco.

## Sobre o Conjunto de Dados

O dataset traz várias informações importantes: idade, IMC, colesterol, triglicerídeos, ureia, creatinina e a Hemoglobina Glicada (HbA1c), que é um dos marcadores mais utilizados para diagnosticar o diabetes. A variável que indica a condição dos pacientes se chama **CLASS**, e separa eles em três grupos: Não Diabéticos (N), Pré-Diabéticos (P) e Diabéticos (Y).

## Primeiras Observações Visuais

Logo nos primeiros gráficos, a HbA1c se destacou como o melhor indicativo para diferenciar os três grupos. Pacientes diabéticos tinham níveis bem mais altos, com separação visual bem clara nos boxplots.

O IMC também apareceu como um fator de risco. Apesar de existir sobreposição entre as classes, os pacientes diabéticos geralmente apresentaram IMC mais alto, o que sugere uma ligação com sobrepeso.

Quando combinei idade e IMC, a tendência ficou ainda mais forte: boa parte dos diabéticos tinha mais de 50 anos e IMC acima de 25. Ou seja, obesidade somada ao envelhecimento parece aumentar bastante o risco.

## Análises de Correlação

Ao observar a matriz de correlação, triglicerídeos (TG) e VLDL mostraram estar muito conectados. Isso indica que essas duas variáveis trazem quase a mesma informação — algo que pode ser importante na hora de criar modelos preditivos.

Já o colesterol total, apesar de ser um indicador clássico, não mostrou muita diferença entre os grupos. Nesse caso, ele não parece ser um bom marcador isolado para o diagnóstico.

## Principais Insights

A HbA1c se mostrou essencial para monitorar e classificar o estágio glicêmico de um paciente. É uma variável forte e confiável.

Idade e IMC também se destacaram, principalmente quando analisados juntos. Eles ajudam a montar um perfil mais claro dos pacientes com maior risco.

Algumas variáveis estavam muito correlacionadas entre si, como TG e VLDL. Isso mostra a importância de escolher bem os atributos na hora de criar modelos, para evitar repetir informações.

Em contrapartida, variáveis como ureia e creatinina — relacionadas à função renal — não se destacaram muito visualmente nesse conjunto.

## **Conclusão**

A análise ajudou a entender melhor o perfil de quem tem diabetes e quais fatores parecem estar mais ligados à condição. A HbA1c, o IMC e a idade se mostraram indicadores importantes, enquanto outras variáveis, como TG e VLDL, levantam um alerta sobre redundância. Com essas descobertas, dá pra pensar em estratégias melhores tanto para diagnóstico quanto para modelos preditivos — além de ajudar na criação de ações de prevenção mais eficientes.