

# 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, como idade, IMC, colesterol, triglicerídeos, uréia, 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 os registros 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 apresentaram níveis bem mais altos, com uma separação visual clara nos boxplots.

O IMC também apareceu como um fator de risco. Apesar de existir certa sobreposição entre as classes, os pacientes diabéticos costumam ter IMC mais elevado, sugerindo uma possível ligação com o sobrepeso.

Quando combinei idade e IMC, o padrão ficou ainda mais evidente. Boa parte dos pacientes diabéticos tinha mais de 50 anos e IMC acima de 25. Ou seja, a combinação entre envelhecimento e obesidade parece estar diretamente ligada ao risco de diabetes.

## Análises de Correlação

A matriz de correlação revelou que triglicerídeos (TG) e VLDL estão altamente conectados. Isso indica que essas variáveis carregam informações muito semelhantes e, em modelos preditivos, talvez não precisem ser usadas juntas.

Já o colesterol total, por outro lado, não apresentou uma boa separação entre os grupos. Isso sugere que, de forma isolada, ele pode não ser um bom marcador para diagnóstico da diabetes neste conjunto de dados.

## **Principais Insights**

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

A idade e o IMC também se destacaram, especialmente quando analisados em conjunto. Eles ajudam a formar um perfil mais preciso dos pacientes com maior risco de desenvolver diabetes.

Variáveis altamente correlacionadas, como TG e VLDL, reforçam a importância de fazer uma boa seleção de atributos em futuros modelos, evitando redundância e melhorando a eficiência da análise.

Já variáveis como ureia e creatinina, ligadas à função renal, não tiveram destaque visual relevante nesse conjunto específico.

## **Conclusão**

Essa análise exploratória ajudou a entender melhor o perfil dos pacientes com diabetes e os fatores mais associados à condição. A Hemoglobina Glicada (HbA1c), o IMC e a idade se mostraram variáveis importantes para se observar com atenção. Ao mesmo tempo, o estudo reforça a importância de analisar possíveis redundâncias nos dados e pensar em estratégias mais inteligentes tanto para diagnóstico quanto para prevenção e tratamento.