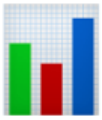


Guia Explicativo — Diagnóstico de Diabetes

1. Introdução ao problema

O diabetes mellitus tipo 2 é uma condição crônica e silenciosa que pode gerar graves complicações à saúde, como doenças cardiovasculares, cegueira e falência renal. Identificar precocemente indivíduos com alto risco de desenvolver diabetes é crucial para intervenção preventiva. Este trabalho busca desenvolver modelos de aprendizado de máquina capazes de prever a presença de diabetes a partir de dados clínicos e laboratoriais de pacientes.

---

2. Descrição do dataset

Foi utilizado o dataset *Pima Indians Diabetes*, composto por 768 registros de mulheres com pelo menos 21 anos de idade e descendência Pima. As colunas são:

- **Gravidezes**: número de gestações
- **Glicose**: concentração de glicose plasmática
- **PressaoArterial**: pressão arterial diastólica
- **EspessuraDobraCutanea**: espessura da dobra cutânea do tríceps
- **Insulina**: insulina sérica
- **IMC**: índice de massa corporal
- **HistoricoFamiliar**: função da história familiar de diabetes
- **Idade**: idade em anos
- **Diabetes**: variável-alvo (0 = não, 1 = sim)

---

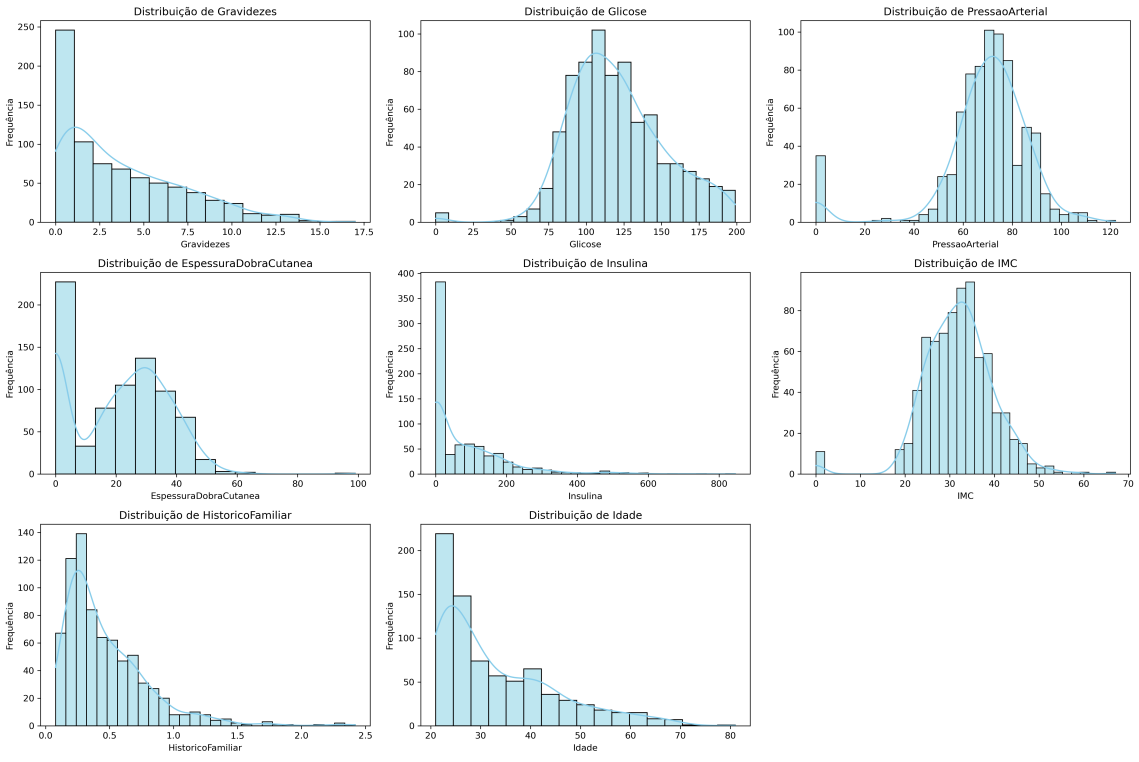
3. EDA e preparação dos dados

📌 Etapas realizadas:

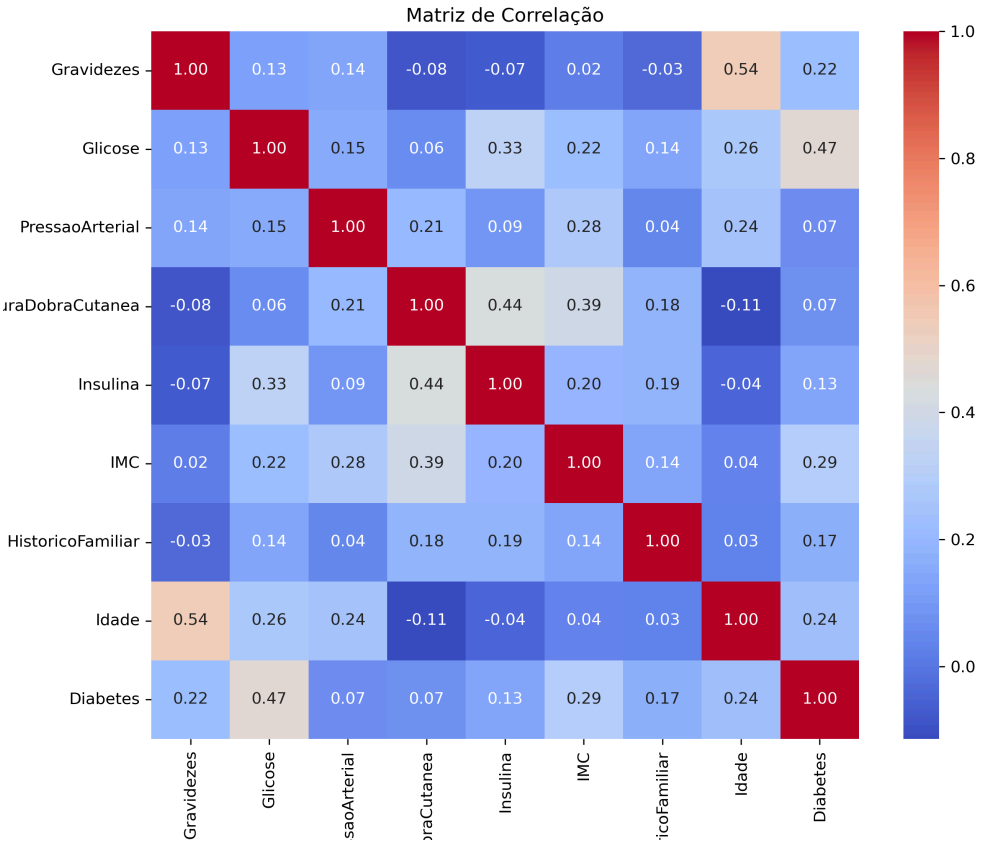
- **Renomeação das colunas** para o português.
- **Análise de estatísticas básicas** e visualizações para entender a distribuição das variáveis.
- **Detecção de valores inválidos** (como glicose = 0, que é clinicamente impossível).
- **Substituição dos zeros por** `NaN` e preenchimento com a **mediana** da respectiva variável.
- **Criação da variável** `FaixaEtaria`, com categorias: Jovem, Adulto, Idoso.
- **Normalização dos dados contínuos** com Z-score (média 0, desvio padrão 1).

📈 Visualizações geradas:

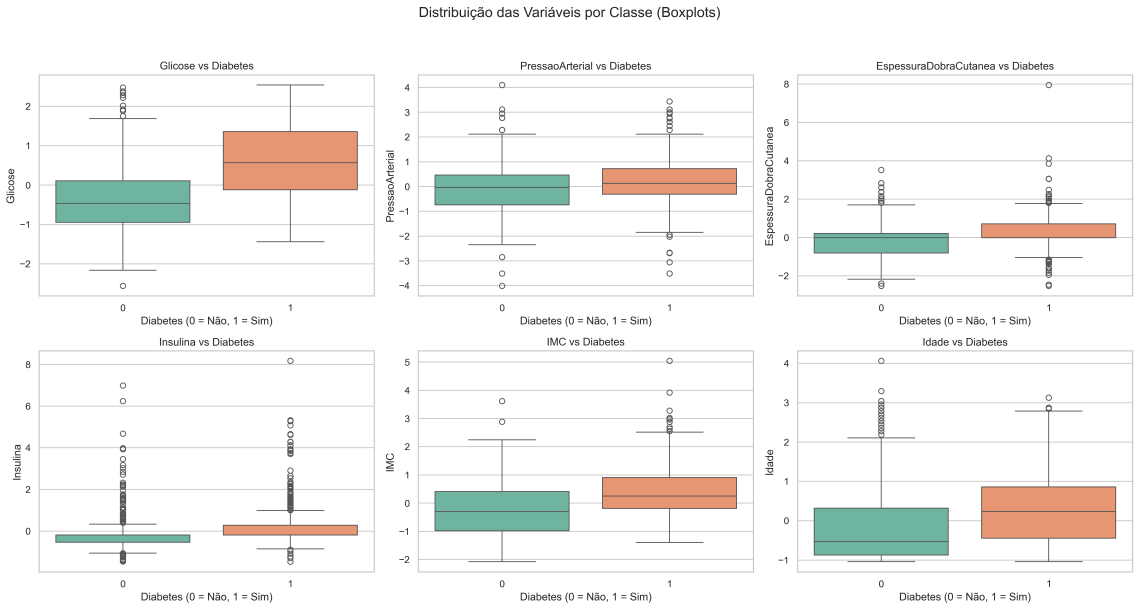
- Distribuição das variáveis contínuas antes da normalização.



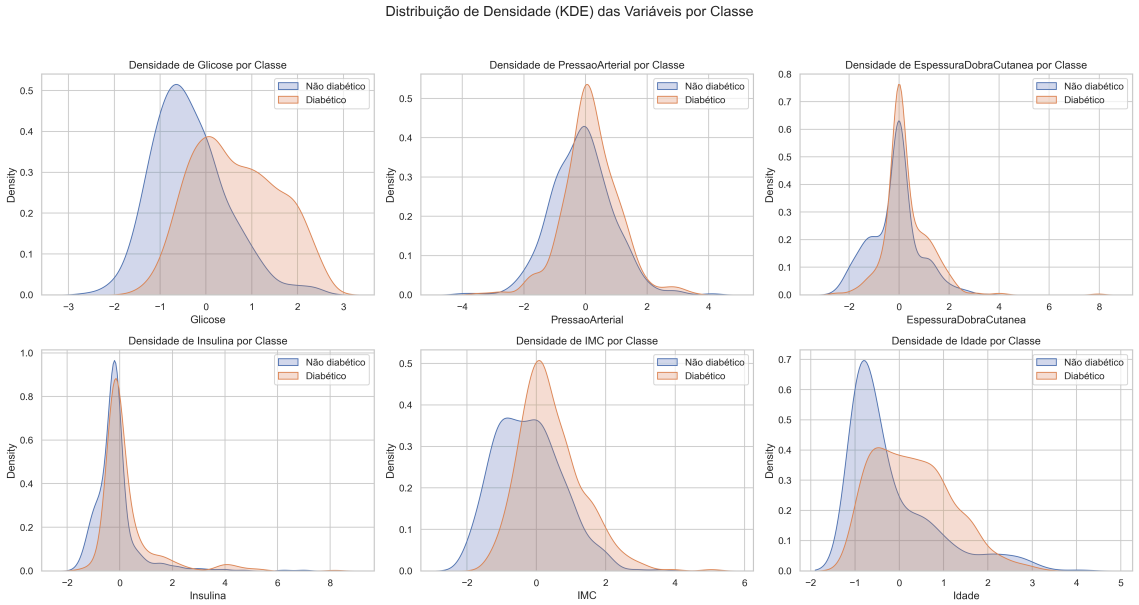
- Matriz de correlação entre as variáveis do dataset.



- Boxplots das variáveis contínuas por classe (0 = não diabético, 1 = diabético).



- Gráficos de densidade (KDE) das variáveis por classe.



- Histogramas por variável
- Matriz de correlação
- Comparações antes/depois da normalização
- Boxplots por classe (diabético/não diabético)
- Gráficos de densidade (KDE) comparando distribuições por classe

---

4. Descrição dos modelos implementados

Dois modelos supervisionados foram utilizados para previsão da variável `Diabetes`:

🔁 Random Forest

- Algoritmo baseado em múltiplas árvores de decisão.
- Vantagens: robustez a overfitting, lida bem com dados com ruído.
- Foi avaliado com métricas como Acurácia, Precisão, Recall e F1-Score.

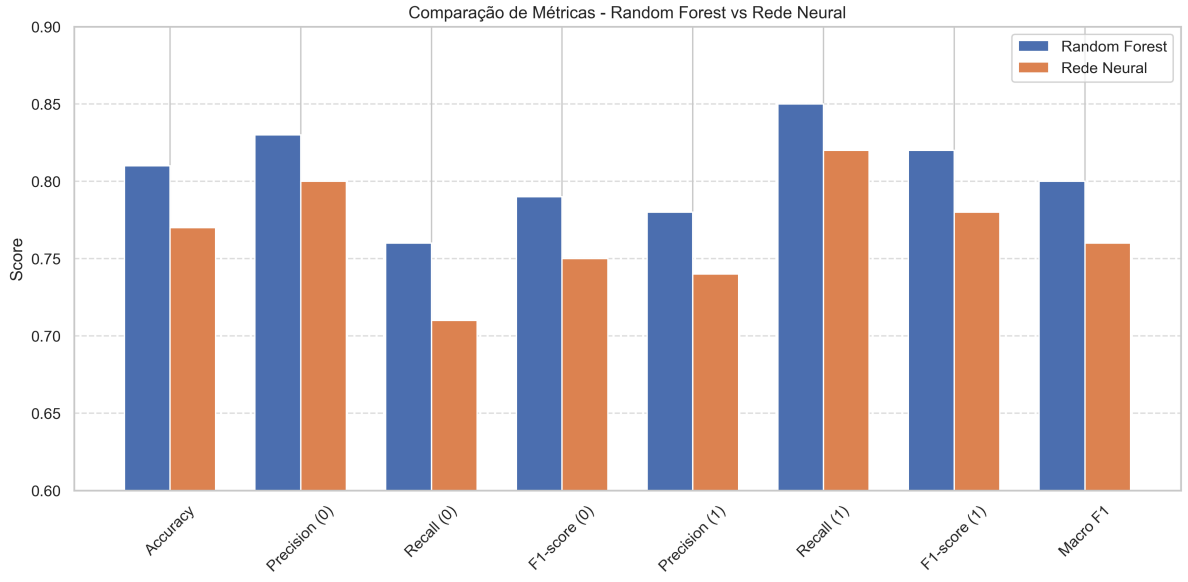
🧠 Rede Neural Artificial (RNA)

- Modelo com camadas densas totalmente conectadas.
- Capaz de capturar relações não lineares e complexas.
- Também avaliado com as mesmas métricas de classificação.

---

5. Resultados e comparação entre modelos

Métrica	Random Forest	Rede Neural
Acurácia	0.79	0.77
Precisão	0.78	0.80
Recall	0.85	0.71
F1-Score	0.82	0.75
ROC-AUC	0.80	0.74



📊 Observações:

- O **Random Forest** teve melhor desempenho geral em quase todas as métricas, especialmente em Recall e F1.
- A **Rede Neural** apresentou bom desempenho em precisão, mas teve recall inferior — indicando que errou mais nos casos positivos (diabéticos).
- As visualizações ajudaram a entender a separação das classes por variáveis, destacando a importância de atributos como Glicose, IMC e Idade.

---

6. Conclusões finais com aprendizados

- A limpeza e o preparo correto dos dados foram fundamentais para garantir a qualidade dos modelos.

- A visualização de dados permitiu insights importantes sobre a relação entre variáveis e o diagnóstico de diabetes.
- O **modelo Random Forest** demonstrou ser uma solução robusta, precisa e interpretável para o problema.
- Aprendemos que diferentes modelos têm pontos fortes e fracos, e que a escolha ideal depende do equilíbrio entre precisão e sensibilidade.
- A inclusão de variáveis derivadas (como `FaixaEtaria`) pode enriquecer os dados e melhorar o desempenho preditivo.
- A normalização foi especialmente relevante para algoritmos baseados em distância ou gradientes.