

**APLICAÇÃO DE MACHINE LEARNING NA  
IDENTIFICAÇÃO DE PADRÕES DE SAÚDE:  
DIAGNÓSTICO DE DIABETES**

**FILIPE OLIVEIRA**

**SÃO PAULO**

**2025**

# SUMÁRIO

1. INTRODUÇÃO	3
2. EXPLORAÇÃO DOS DADOS (EDA)	3
3. PRÉ-PROCESSAMENTO	5
3.1 <i>Tratamento de valores inválidos</i>	6
3.2 <i>Normalização das variáveis</i>	7
3.3 <i>Balanceamento das classes</i>	9
4. AVALIAÇÃO E COMPARAÇÃO DOS MODELOS	9
5. INTERPRETAÇÃO COM TÉCNICAS DE EXPLICABILIDADE	10
7. DISCUSSÃO CRÍTICA	11
8. CONCLUSÃO	12

## 1. Introdução

O diagnóstico precoce do diabetes mellitus é um dos principais desafios enfrentados pelos sistemas de saúde atualmente. Trata-se de uma condição crônica que pode causar complicações graves se não for identificada e controlada a tempo. Nesse contexto, modelos de aprendizado de máquina podem auxiliar na identificação de padrões em dados clínicos, apoiando profissionais da saúde na triagem e detecção de pacientes com maior probabilidade de desenvolver a doença.

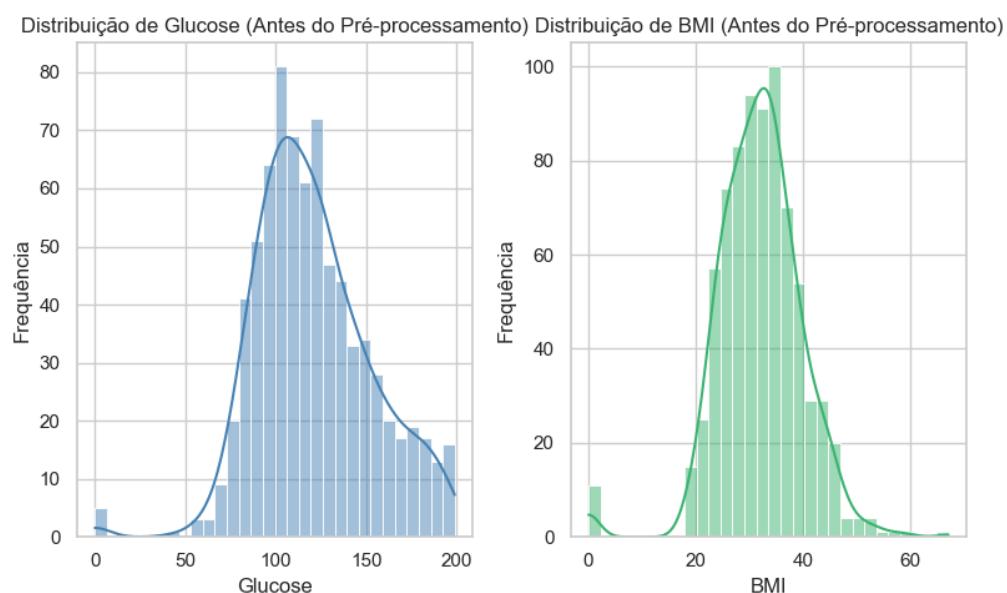
Para este estudo, foi utilizado um conjunto de dados público amplamente referenciado na literatura, contendo informações clínicas como glicose, índice de massa corporal (BMI), número de gestações, entre outros fatores relacionados ao risco de diabetes. A escolha desse dataset se deu por sua disponibilidade e simplicidade de aplicação, permitindo concentrar o foco nas etapas de pré-processamento, modelagem e análise de resultados.

O conjunto de dados foi obtido na plataforma [Kaggle](#), e encontra-se disponível para consulta no repositório do projeto que pode ser encontrado no [GitHub](#).

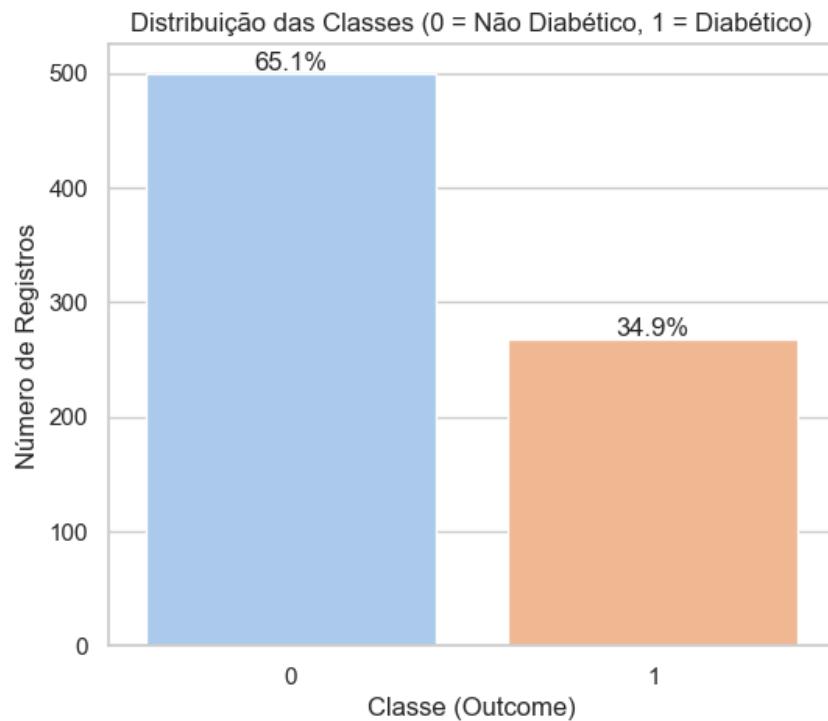
## 2. Exploração dos dados (EDA)

A etapa de exploração dos dados teve como objetivo compreender a estrutura do conjunto e identificar possíveis problemas de qualidade que pudessem impactar o desempenho dos modelos preditivos.

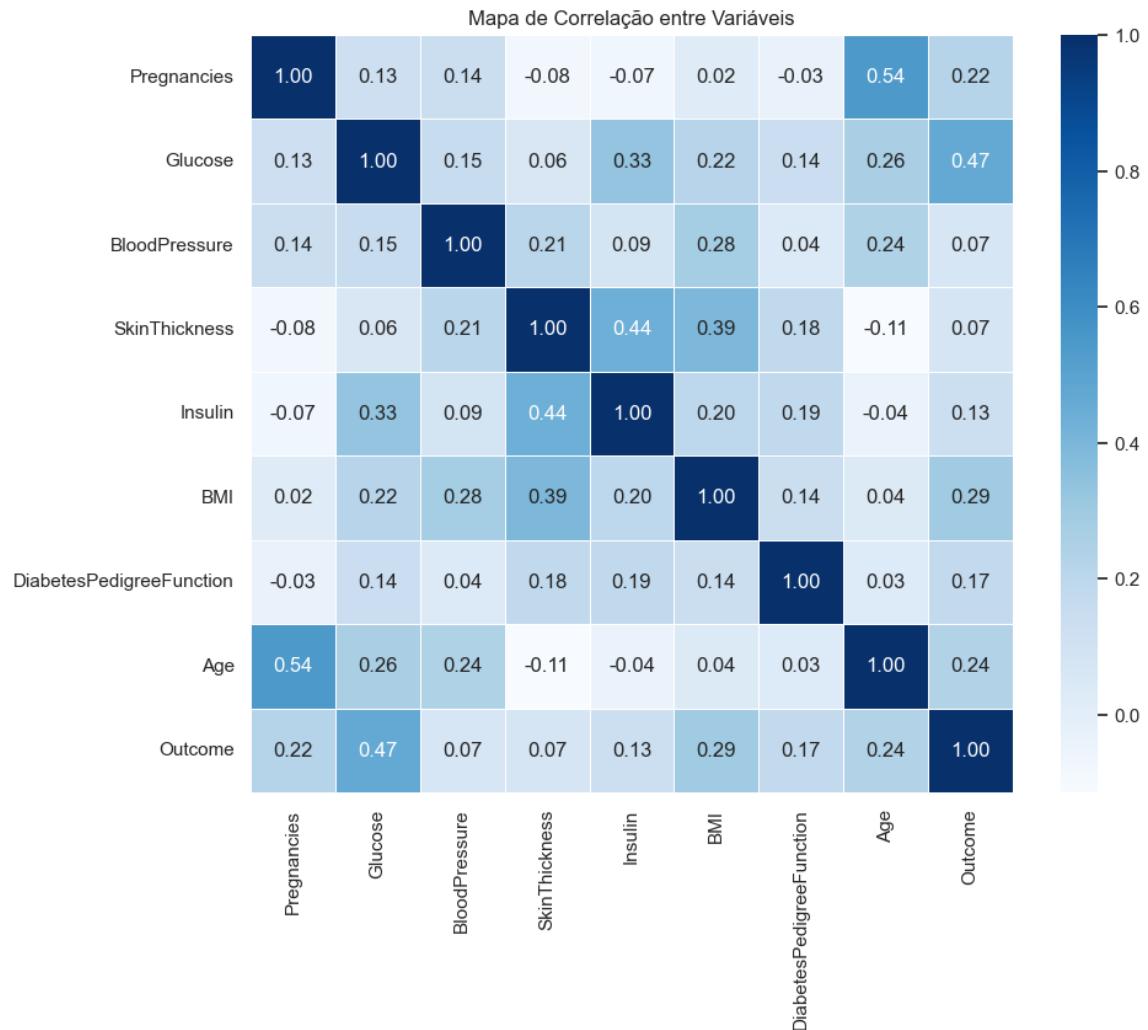
Durante a análise inicial, observou-se que algumas variáveis continham **valores igual a zero**, o que é fisiologicamente impossível em medições médicas. Esse problema foi identificado nas colunas **Glucose**, **BloodPressure**, **SkinThickness**, **Insulin** e **BMI**. Para corrigir essa inconsistência, esses valores foram substituídos por *Nan* e posteriormente tratados por meio de **imputação com o algoritmo KNNImputer**, garantindo maior consistência ao conjunto de dados.



Ao analisar a **distribuição das classes**, verificou-se que aproximadamente **65% dos registros pertencem à classe 0 (não diabético) e 35% à classe 1 (diabético)**. Esse leve desequilíbrio poderia influenciar o aprendizado do modelo, levando-o a favorecer a classe majoritária. Assim, decidiu-se aplicar a técnica **SMOTE (Synthetic Minority Over-sampling Technique)** após a divisão treino/teste, para balancear as classes e melhorar a capacidade de generalização do modelo.



A análise de correlação revelou que as variáveis **Glucose** e **BMI** possuem as maiores associações com o resultado **Outcome**, indicando forte relação com a presença de diabetes. Já as variáveis **SkinThickness** e **Insulin** apresentaram correlação mais baixa, sugerindo menor influência direta sobre o diagnóstico dentro deste conjunto de dados.



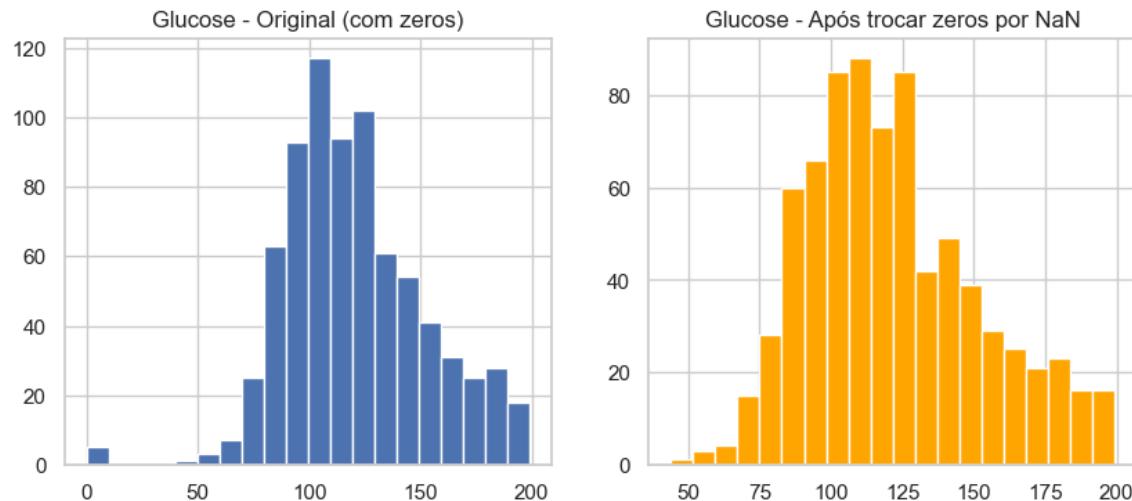
Esses achados orientaram as etapas seguintes de **pré-processamento e modelagem**, garantindo que o pipeline considerasse tanto a limpeza quanto o balanceamento adequado dos dados antes do treinamento dos modelos.

### 3. Pré-processamento

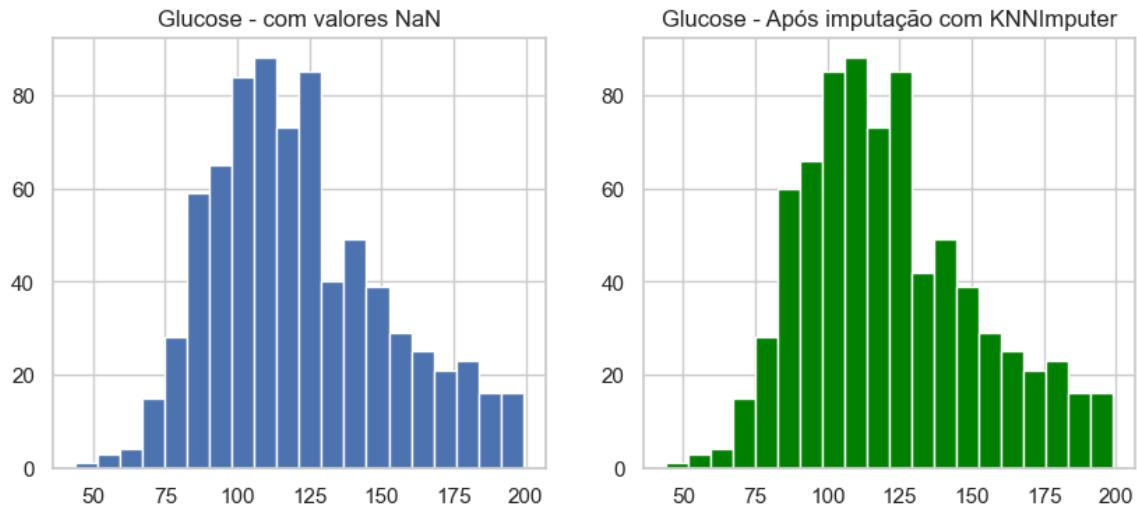
O pré-processamento teve como objetivo preparar o conjunto de dados para a modelagem, garantindo consistência e qualidade nas informações utilizadas pelos algoritmos. As etapas realizadas incluíram o tratamento de valores inválidos, a imputação de dados ausentes, a normalização das variáveis e o balanceamento das classes.

### 3.1 Tratamento de valores inválidos

Inicialmente, foi identificado que algumas colunas — *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin* e *BMI* — continham valores iguais a zero, que não são fisiologicamente possíveis e representam dados ausentes. Esses valores foram substituídos por valores nulos (*NaN*).



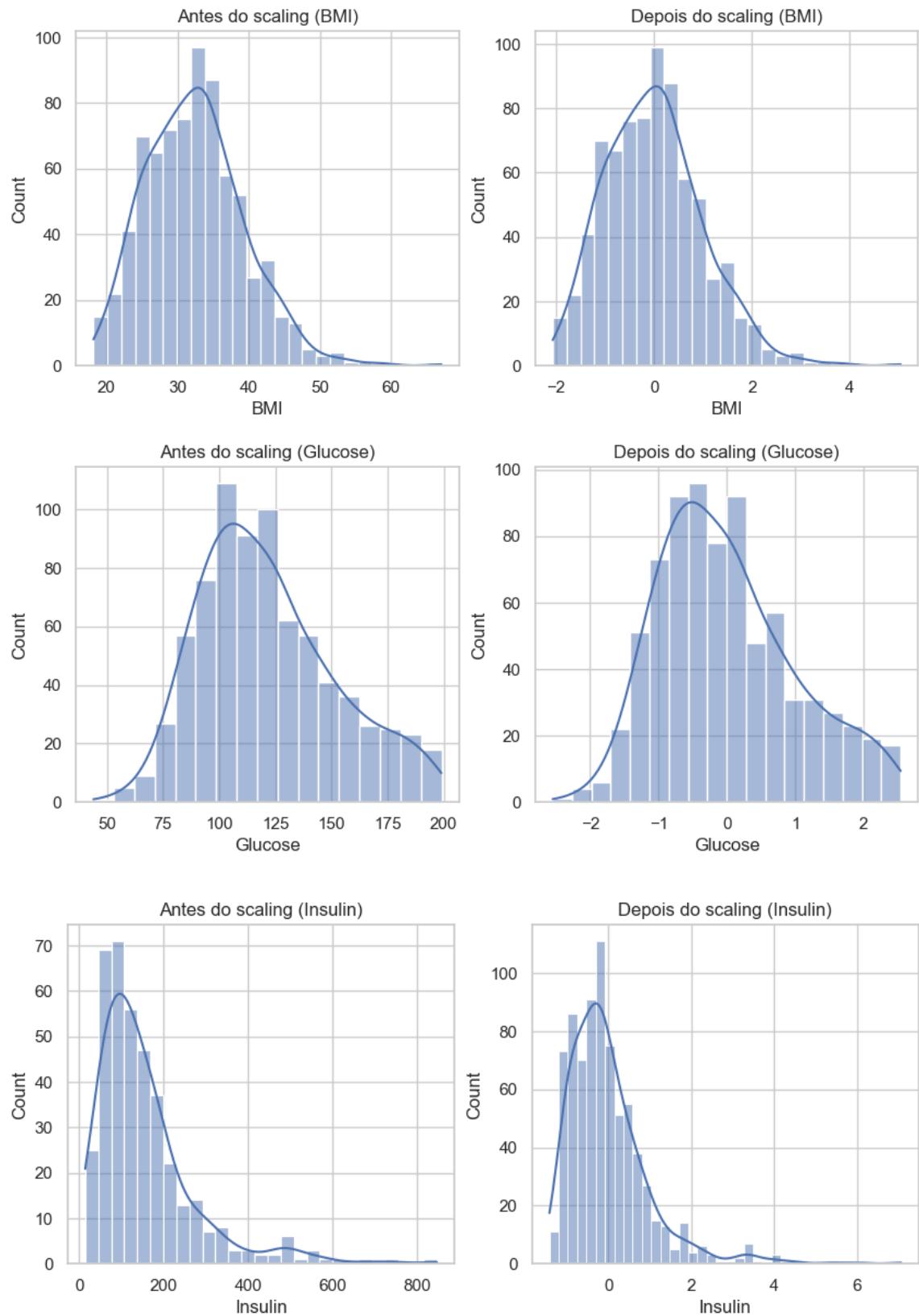
Em seguida, aplicou-se o método **KNNImputer** para preencher os valores ausentes com base na similaridade entre os registros do conjunto de dados.



Após essa etapa, observou-se que a distribuição das variáveis permaneceu coerente, sem introdução de outliers artificiais, o que indica que o processo preservou as características originais do conjunto.

### 3.2 Normalização das variáveis

Na sequência, aplicou-se o **StandardScaler** para normalizar os dados numéricos, ajustando todas as variáveis para média zero e desvio padrão igual a um. Essa padronização garante que atributos com escalas diferentes — como *Insulin* e *BloodPressure* — não dominem o processo de aprendizado dos modelos.

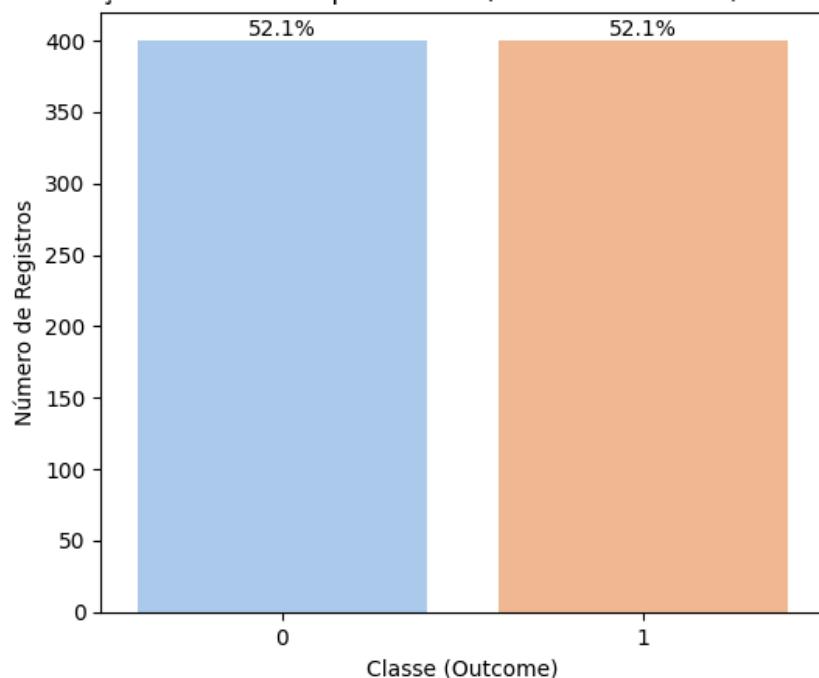


Esses gráficos evidenciam que, após o *scaling*, as distribuições permanecem semelhantes, porém ajustadas em escala, o que favorece o desempenho dos algoritmos baseados em distância, como regressão logística e KNN.

### 3.3 Balanceamento das classes

Por fim, considerando que o conjunto apresentava um leve desbalanceamento entre as classes da variável alvo (*Outcome*), foi utilizada a técnica **SMOTE (Synthetic Minority Over-sampling Technique)** para gerar amostras sintéticas da classe minoritária. Essa abordagem contribui para que o modelo aprenda de forma equilibrada, aumentando sua capacidade de identificar corretamente casos positivos de diabetes.

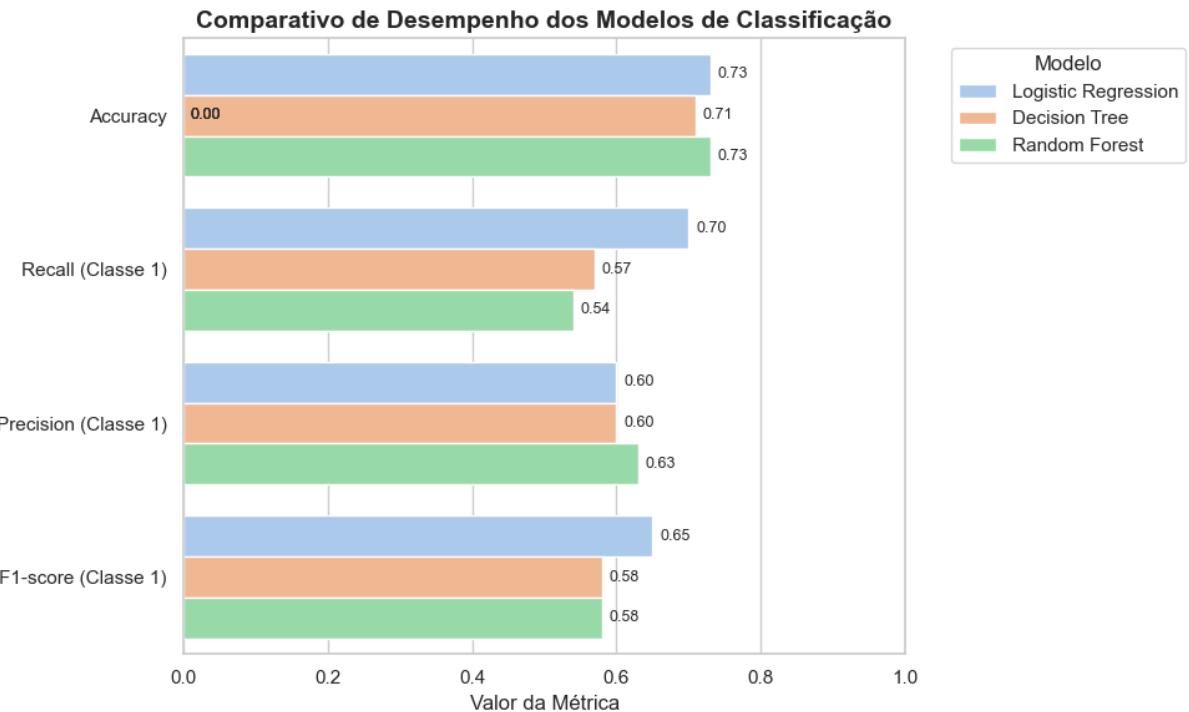
Distribuição das Classes Após SMOTE (0 = Não Diabético, 1 = Diabético)



## 4. Avaliação e Comparação dos Modelos

No projeto foram utilizados **três modelos de classificação: Regressão Logística, Decision Tree e Random Forest**. Cada um deles foi selecionado por características complementares, permitindo comparar desempenho e interpretar resultados de diferentes perspectivas:

- **Regressão Logística (Logistic Regression)** – Modelo Linear
- **Decision Tree (Árvore de Decisão)** - Modelo não linear
- **Random Forest** - Conjunto de árvores de decisão



O modelo de Regressão Logística apresentou o maior recall (0.70), o que significa que ele foi o mais eficaz em detectar corretamente pacientes com a doença.

Embora o modelo Random Forest tenha alcançado uma precisão ligeiramente maior (0.63), seu recall (0.54) foi inferior, o que implica maior risco de falsos negativos — um problema crítico em diagnósticos médicos.

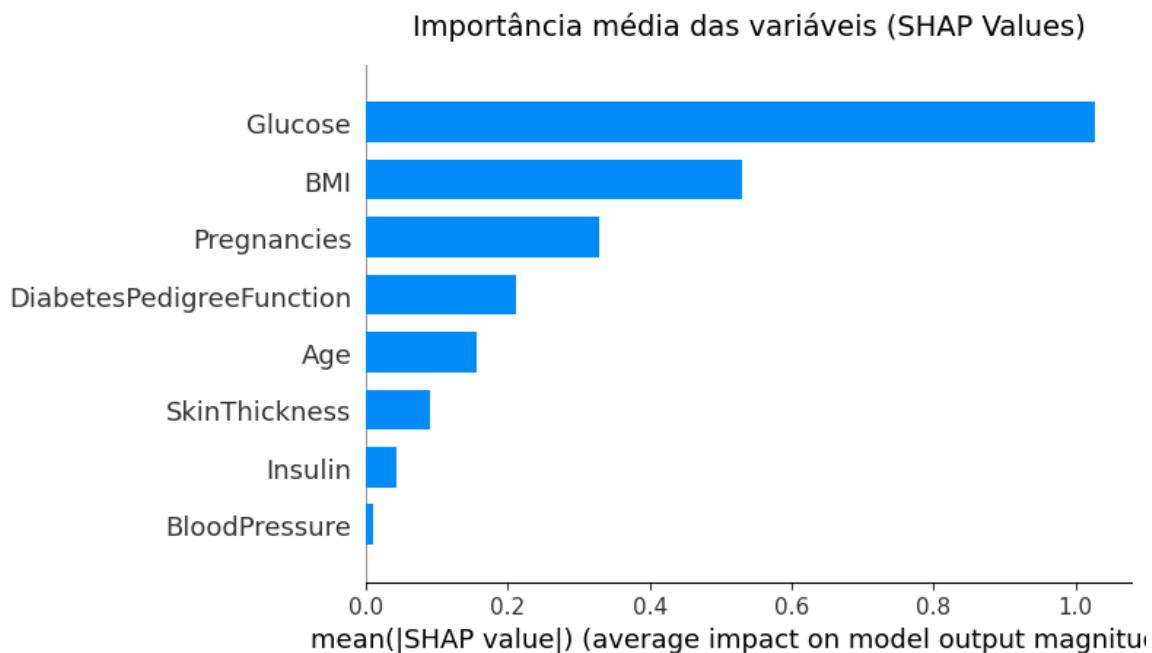
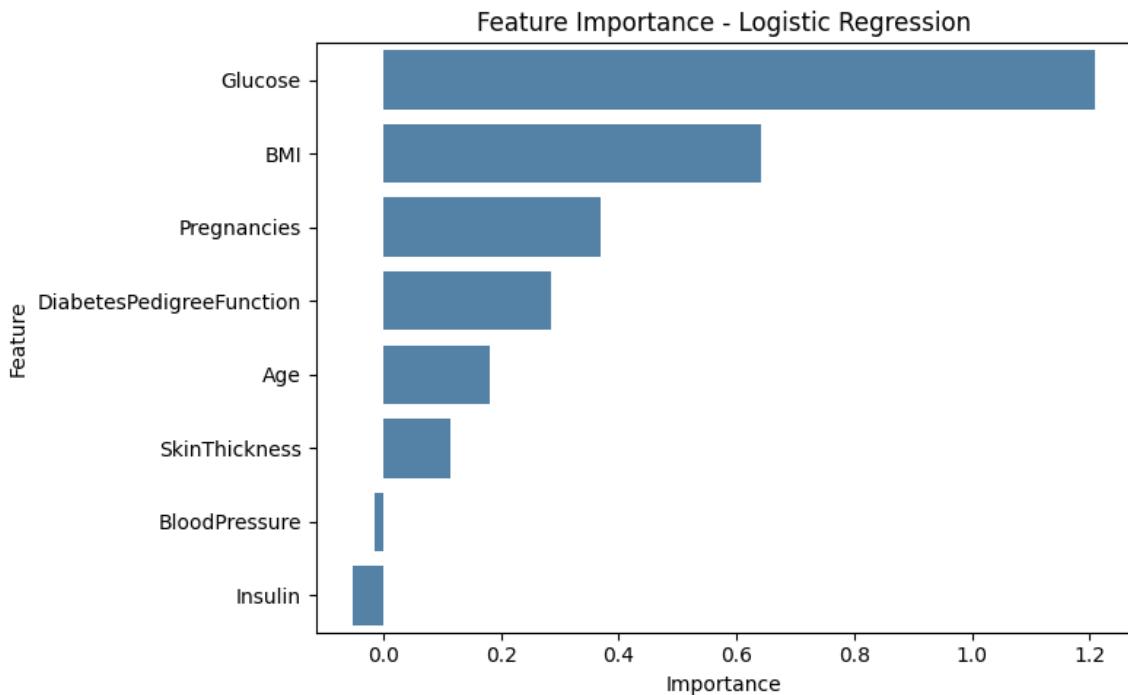
O modelo de Árvore de Decisão apresentou desempenho equilibrado, mas sem destaque em nenhuma métrica específica.

Como em cenários clínicos a prioridade é minimizar falsos negativos (evitar deixar de identificar pacientes realmente doentes), o modelo Logistic Regression foi selecionado como o mais adequado para este pipeline.

## 5. Interpretação com Técnicas de Explicabilidade

Para compreender as variáveis que mais influenciaram as previsões do modelo final (Regressão Logística), foram aplicadas as técnicas de **Feature Importance** e **SHAP (SHapley Additive Explanations)**.

Ambas as abordagens destacaram **Glucose**, **BMI** e **Pregnancies** como as variáveis mais determinantes para o diagnóstico de diabetes. A análise de SHAP, representada no gráfico de barras a seguir, confirma que valores mais altos dessas variáveis tendem a aumentar a probabilidade do modelo prever a presença da doença.



## 7. Discussão Crítica

Embora o modelo de **Regressão Logística** tenha apresentado bom desempenho (*recall* = 0.70 e *acurácia* = 0.73), é importante reconhecer suas limitações:

- O **dataset** utilizado é limitado em tamanho e diversidade, o que pode afetar a capacidade de generalização do modelo para novas populações.
- Existe **desbalanceamento entre classes**, o que reduz a capacidade do modelo de detectar corretamente casos da classe minoritária (pacientes diabéticos).
- O modelo **não substitui o julgamento clínico**: deve ser utilizado apenas como ferramenta de apoio à decisão médica, auxiliando o profissional na triagem e priorização de pacientes em risco.

A análise de **Feature Importance** e **SHAP** reforça que variáveis como **Glucose**, **BMI** e **Pregnancies** são determinantes nas predições. Apesar disso, a interpretação das features não elimina a necessidade de supervisão clínica.

Em resumo, o modelo é adequado **como apoio ao diagnóstico**, mas **não deve ser usado de forma autônoma** em contextos clínicos.

## 8. Conclusão

O projeto apresentou um pipeline completo de aprendizado de máquina voltado ao diagnóstico de diabetes, abrangendo desde a exploração e tratamento dos dados até a modelagem e interpretação dos resultados.

Durante a **análise exploratória (EDA)**, foram identificados valores fisiologicamente impossíveis em variáveis como *Glucose* e *BMI*, posteriormente tratados por meio de imputação com **KNNImputer**, garantindo maior consistência aos dados. Além disso, constatou-se um **desequilíbrio moderado** entre as classes, corrigido com o uso de **SMOTE**, de modo a melhorar a capacidade preditiva do modelo sobre casos positivos.

Na etapa de **modelagem**, diferentes algoritmos de classificação foram testados — *Logistic Regression*, *Decision Tree* e *Random Forest*. O modelo de **Regressão Logística** apresentou o melhor desempenho em termos de *recall* (0.70) e *acurácia* (0.73), sendo, portanto, escolhido como o modelo final por priorizar a **sensibilidade** — aspecto essencial em contextos médicos, onde falsos negativos podem ter impacto grave.

A **análise de interpretabilidade** com *Feature Importance* e *SHAP* evidenciou que as variáveis **Glucose**, **BMI** e **Pregnancies** possuem maior influência na predição do desfecho, alinhando-se ao conhecimento médico existente sobre fatores de risco para diabetes.

Por fim, apesar do bom desempenho, o modelo apresenta **limitações relacionadas ao tamanho e à diversidade do conjunto de dados**, o que pode comprometer sua generalização em contextos reais. Assim, recomenda-se seu uso **apenas como ferramenta de apoio** à decisão clínica, servindo para triagem e priorização de pacientes em risco, mas **nunca como substituto da avaliação médica**.

Em síntese, o estudo demonstra o potencial do uso de técnicas de *Machine Learning* na área da saúde, reforçando a importância do uso ético, interpretável e supervisionado dessas soluções em apoio ao diagnóstico médico.