



Análise do risco de diabetes tipo 2 em mulheres Pima (Akimel O'odham) a partir de dados clínicos históricos

Bruno Cerqueira Gianotti | 10721759

Daniel Fernandes Saraiva | 10381985

Gabrielle Solange Ferreira | 10414956

Reginaldo Rogério de Campos | 10743942

Sumário

Sumário.....	2
Introdução.....	3
Premissas do projeto.....	5
Objetivo Geral.....	6
Objetivos Específicos.....	6
Referências de aquisição do dataset.....	7
Descrição da origem dos dados.....	8
Apresentação da Empresa.....	9
Problemas do Estudo.....	14
Pensamento Computacional aplicado à EDA.....	17
Metadados (Descrição do dataset).....	20
Análise Exploratória dos Dados.....	23
Cronograma.....	43
Referências Bibliográficas.....	44
Glossário.....	47

Introdução

A diabetes mellitus tipo 2 é uma das principais doenças crônicas não transmissíveis no mundo, associada a altos índices de morbidade, mortalidade e custos em saúde pública. Entre as populações indígenas da América do Norte, destaca-se a comunidade Pima (Akimel O'odham), localizada no estado do Arizona (EUA), que apresenta uma das maiores prevalências de diabetes tipo 2 já registradas globalmente.

Desde 1965, o National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) tem conduzido um estudo longitudinal sobre essa população, com o objetivo de compreender os fatores clínicos, genéticos e de estilo de vida que contribuem para o desenvolvimento da doença. Como fruto desse esforço, surgiu o Pima Indians Diabetes Dataset, que reúne informações de mulheres com 21 anos ou mais, contemplando variáveis como glicose, pressão arterial, índice de massa corporal, histórico de gravidez, entre outras.

A importância desse conjunto de dados vai além da análise epidemiológica: ele tornou-se um dos datasets mais utilizados em ciência de dados e aprendizado de máquina, servindo como referência para o desenvolvimento e validação de modelos preditivos em saúde. Assim, ele contribui não apenas para avanços acadêmicos, mas também para a formulação de estratégias de

prevenção, diagnóstico precoce e políticas de saúde voltadas a populações vulneráveis.

Nesse contexto, este trabalho tem como objetivo analisar o Pima Indians Diabetes Database, disponibilizado atualmente em repositórios como o Kaggle, buscando compreender como variáveis clínicas e históricas se relacionam com o diagnóstico de diabetes tipo 2. O estudo também representa uma oportunidade de aplicar conceitos de ciência de dados, explorando padrões, correlações e potenciais fatores de risco.

Todo o código, dataset e versões deste projeto estão disponíveis no repositório GitHub:

<https://github.com/BrunoCerqueiraGianotti/projeto-diabetes-pima>.

Premissas do projeto

Diante dessa relevância, este projeto parte do princípio de que a análise de dados clínicos e sociodemográficos pode fornecer insights valiosos para a predição e compreensão da diabetes tipo 2.

A premissa central consiste em que, por meio de técnicas de ciência de dados, é possível identificar padrões de risco, avaliar os fatores mais influentes no desenvolvimento da doença e construir modelos preditivos capazes de apoiar estratégias de prevenção e diagnóstico precoce.

Objetivo Geral

Desenvolver um projeto aplicado de ciência de dados utilizando o Pima Indians Diabetes Dataset, com o propósito de analisar fatores de risco através da análise exploratória de dados e que permita estimar a probabilidade de ocorrência de diabetes tipo 2, contribuindo para a compreensão da doença e para estratégias de prevenção.

Objetivos Específicos

1. Realizar análise exploratória dos dados (EDA) para compreender a distribuição das variáveis, identificar outliers e tratar dados faltantes.
2. Investigar a correlação entre fatores clínicos/demográficos e a presença de diabetes, destacando os mais relevantes.
3. Aplicar técnicas de pré-processamento, como normalização e balanceamento de classes, para preparar os dados para modelagem.
4. Interpretar os resultados obtidos e discutir suas implicações para o contexto de saúde pública da população Pima.

Referências de aquisição do dataset

Os dados foram coletados por meio de exames clínicos e entrevistas estruturadas conduzidas pela equipe do NIDDK, como parte de um estudo epidemiológico de longo prazo com foco na diabetes tipo 2. A população estudada consiste exclusivamente em mulheres da comunidade Pima, acima de 21 anos, localizadas no sul do Arizona. O contexto envolve variáveis relacionadas à saúde metabólica, fatores de risco e histórico familiar, com o objetivo de rastrear padrões emergentes e fatores predisponentes.

Período da coleta: desde 1965, em ciclos bienais de exames clínicos e entrevistas.

População: mulheres de ascendência Pima, residentes no Arizona, com idade ≥ 21 anos.

Limitações: presença de valores zero em variáveis que não poderiam assumir esse valor (e.g., glicose, IMC), indicando a necessidade de pré-processamento.

Licenciamento e uso: o dataset é amplamente utilizado em pesquisas e práticas de ensino, servindo como benchmark em algoritmos de classificação e aprendizado supervisionado.

Descrição da origem dos dados

A origem dos dados é um estudo epidemiológico de longo prazo conduzido pelo NIDDK. O foco principal foi monitorar a alta incidência de diabetes tipo 2 entre mulheres Pima, relacionando-a a fatores metabólicos, clínicos e familiares.

O levantamento foi realizado com:

Exames laboratoriais: medições de glicose, insulina e IMC.

Entrevistas estruturadas: histórico de saúde e condições familiares.

Variáveis antropométricas: idade, número de gestações, espessura da pele, pressão arterial.

Esse contexto justifica a relevância científica do dataset, que não só contribui para pesquisas médicas, como também é amplamente utilizado como referência em projetos de ciência de dados e aprendizado de máquina.

Apresentação da Empresa

Nome da empresa : O National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)

Apresentação da Empresa: O NIDDK é uma agência governamental dos Estados Unidos da América, parte do National Institutes of Health (NIH), a principal agência de pesquisa médica do país, que por sua vez pertence ao Departamento de Saúde e Serviços Humanos (Department of Health and Human Services). O Instituto financia uma vasta gama de pesquisas médicas por meio de bolsas a universidades e outras instituições, e também mantém cientistas governamentais que conduzem pesquisas básicas, translacionais e clínicas em seus laboratórios internos.

Missão/Visão/Valores:

- **Missão:** Conduzir e apoiar pesquisas médicas e treinamento em pesquisa e disseminar informações baseadas em ciência sobre diabetes e outras doenças endócrinas e metabólicas; doenças digestivas, distúrbios nutricionais e obesidade; e doenças renais, urológicas e hematológicas, para melhorar a saúde e a qualidade de vida das pessoas.
- **Visão:** Continuar a buscar as pesquisas mais essenciais para combater as muitas doenças crônicas debilitantes e dispendiosas dentro de sua missão, mantendo um firme compromisso com a pesquisa básica, translacional e clínica; treinamento em pesquisa e desenvolvimento de carreira; e a disseminação de informações de saúde para melhorar a vida

dos pacientes, suas famílias e aqueles em risco dessas doenças.

- **Valores:**

- **Integridade:** Manter padrões de integridade e ética inabaláveis.
- **Colaboração:** Promover um ambiente colaborativo para alcançar objetivos comuns.
- **Respeito:** Tratar cada indivíduo com dignidade, empatia e respeito.
- **Transparência:** Operar com transparência e abertura.

Segmento de atuação: O NIDDK atua no Segmento de Pesquisa Biomédica e Saúde Pública.

Sua importância para o Estado: Como uma agência governamental pública federal dos EUA, o NIDDK não possui **market share**. Sua importância para o Estado (o governo e a nação dos EUA) é descrita por:

- **Liderança em Pesquisa:** É uma das maiores instituições do NIH, dedicando-se a doenças crônicas, complexas e com grandes consequências para a saúde pública (como diabetes, doenças renais e obesidade).
- **Financiamento Vital:** Recebe dotações federais do Congresso para financiar pesquisas em todo o país, sendo crucial para o ecossistema de pesquisa biomédica dos EUA. O orçamento discricionário do NIDDK para o ano fiscal de 2024 foi de mais de \$2.3 bilhões.

- **Impacto na Saúde Pública:** A pesquisa apoiada pelo NIDDK levou a avanços significativos, como o desenvolvimento de insulinas de ação prolongada e medicamentos anti-obesidade, melhorando a saúde e a qualidade de vida de milhões de pessoas. O Instituto também desempenha um papel fundamental na formação de novos investigadores científicos.

A 'concorrência' no âmbito da pesquisa, mesmo que não compita comercialmente, o NIDDK atua em um campo onde outras instituições e grupos de pesquisa buscam os mesmos objetivos científicos e, por vezes, os mesmos recursos (como financiamento e talentos). Essas entidades podem ser vistas como seus "concorrentes" no cenário científico, porém a relação com essas entidades é, em grande parte, de colaboração e parceria, para impulsionar a excelência e o progresso das pesquisas.

Exemplos de Entidades no Mesmo Campo de Atuação:

- **Outros Institutos do NIH** - o NIDDK é um dos 27 institutos do NIH. Outros institutos podem ter linhas de pesquisa que se sobrepõem à sua missão.
- **Universidades e Centros de Pesquisa** - Instituições acadêmicas são parceiras fundamentais em projetos de pesquisa, muitas vezes financiadas pelo próprio NIDDK.
- **Fundações Privadas** - Organizações não governamentais dedicadas a doenças específicas também financiam e realizam pesquisas na mesma área de atuação.
- **Consórcios de Pesquisa** - O próprio NIDDK coordena esforços por meio de consórcios (ex.: George M. O'Brien

Kidney Consortium), demonstrando que a colaboração é a regra.

Número de Colaboradores: O NIDDK gerencia uma equipe de mais de 690 funcionários/servidores públicos (dado de 2024, referente ao staff interno).

Iniciativas na Área de Data Science: O NIDDK possui iniciativas estabelecidas, visando alavancar a Ciência de Dados e a Inteligência Artificial (IA) na pesquisa biomédica:

- **NIDDK Central Repository Resources for Research (NIDDK-CR R4R):** Um repositório central que disponibiliza dados e amostras de estudos clínicos significativos para a comunidade de pesquisa, incentivando o reuso e a descoberta impulsionada por dados.
- **NIDDK Information Network (dkNET):** Um portal que apoia a ciência robusta e reprodutível, ajudando a autenticar recursos e a buscar *datasets* existentes.
- **NIDDK-CR Data Science e Meet the Expert Webinar Series:** Uma série de *webinars* mensais focada em acelerar a Ciência de Dados e a pesquisa biomédica impulsionada por IA, cobrindo o ciclo de vida da Ciência de Dados, metodologias avançadas e fundamentos de IA.

Trabalhos em Destaque (com Ciência de Dados e IA):

- **Desafios de Dados (Data Challenges):** O NIDDK-CR hospeda plataformas de desafios de dados visando aumentar e aprimorar os dados existentes no Repositório para

pesquisas secundárias futuras, incluindo descobertas impulsionadas por IA.

- Um **Data Centric Challenge** foi anunciado em 2023, focado em aprimorar os conjuntos de dados do NIDDK para futuras aplicações de IA e *Machine Learning* (ML).
- **Melhoria da Qualidade dos Dados para a Preparação para IA:** O NIDDK-CR tem implementado esforços desde 2021 para melhorar a qualidade dos dados para que estejam prontos para IA (*AI-readiness*), incluindo o uso de Processamento de Linguagem Natural (NLP) em projetos piloto e a adoção de padrões de dados para alinhamento com os princípios FAIR.

Problemas do Estudo

Limitações e problemas identificados

- Restrição da amostra a uma população específica, o que dificulta a generalização dos resultados para outros grupos étnicos e contextos socioculturais.
- O artigo estabelece associações entre fatores (obesidade, glicose, hereditariedade), mas não demonstra causalidade direta.
- Ausência de identificação dos genes específicos associados ao DM2. Os autores reconhecem que, embora haja clara agregação familiar da doença, não foram localizados os determinantes genéticos exatos — uma limitação compreensível para o período da pesquisa, anterior ao avanço da genômica.
- O estudo também apresenta lacunas quanto aos fatores ambientais, como padrões alimentares, nível de atividade física e condições socioeconômicas. Esses elementos são citados de forma qualitativa, sem medições objetivas, o que impede uma avaliação precisa de seu impacto.
- Os aspectos psicológicos e culturais da comunidade Pima não foram considerados, reduzindo a compreensão dos determinantes sociais da saúde.

O que falta?

O dataset carece de variáveis clínicas complementares, como colesterol, triglicerídeos, histórico familiar detalhado, dieta e nível de

atividade física. Além disso, falta a mensuração da pressão sistólica para uma avaliação cardiovascular completa.

É identificado também a falta do contexto clínico, pois o diabetes é multifatorial, e o dataset foca apenas em medições físicas e laboratoriais simples.

O que incomoda?

- A alta proporção de valores ausentes em Insulin e SkinThickness limita a confiabilidade da análise.
- DiabetesPedigreeFunction — índice calculado artificialmente, com cálculo não transparente e pouca relevância clínica isolada.
- Pressão Arterial Diastólica (BloodPressure) — pouco informativa isoladamente, pois o ideal seria considerar pressão sistólica ou a pressão arterial média.

Há um padrão que pode ser observado?

- Glicose tem a maior correlação com o diagnóstico de diabetes. Clinicamente faz sentido, pois a hiperglicemia é o principal critério diagnóstico.
- IMC e Idade têm relação positiva com o desfecho, indicando que o aumento de peso e idade eleva o risco.
- Insulina e Espessura da Pele apresentam correlação entre si, refletindo adiposidade corporal.
- BloodPressure e DiabetesPedigreeFunction são as variáveis menos correlacionadas com o desfecho.

Há uma afirmação que pode ser contestada?

1. “DiabetesPedigreeFunction é um bom indicador genético de risco.” → Contestável, pois é um índice simplista e não reflete herança genética real.
2. “Pressão arterial diastólica média prediz diabetes.” → Contestável, pois a literatura indica pressão sistólica e média arterial como marcadores mais sensíveis.

Pensamento Computacional aplicado à EDA

A aplicação do Pensamento Computacional (PC) na EDA permite organizar o raciocínio de forma estruturada e sistemática. O estudo do conjunto de dados ‘Pima Indians Diabetes’ foi utilizado como exemplo para demonstrar como os princípios do PC — decomposição, reconhecimento de padrões, abstração e algoritmos — podem ser aplicados ao contexto da Ciência de Dados.

1. Decomposição

Na EDA do *Pima Indians Diabetes Dataset*, o problema de entender os fatores associados ao diabetes foi dividido em etapas sucessivas e interdependentes.

Inicialmente, foram importadas as bibliotecas e carregado os dados, preparando o ambiente de análise. Em seguida, realizou-se a inspeção inicial para compreender a estrutura do dataset e a identificação de valores ausentes, garantindo a qualidade dos dados.

Depois, aplicaram-se medidas estatísticas (de posição e dispersão) e visualizações gráficas (histogramas, boxplots e scatterplots) para explorar o comportamento e a variabilidade das variáveis.

Por fim, foram analisadas as correlações e relações entre variáveis, permitindo compreender os principais padrões clínicos do diabetes.

2. Reconhecimento de Padrões

A análise revelou padrões clínicos esperados no diabetes tipo 2. Observou-se que níveis elevados de glicose, IMC maior e idade

avançada estão associados ao aumento do risco de diabetes, refletindo os principais fatores fisiopatológicos da doença: hiperglicemia, obesidade e resistência à insulina.

Também foi identificada correlação entre Insulina e a espessura da dobra cutânea tricipital, indicando relação com adiposidade corporal. Já a pressão arterial diastólica e a DiabetesPedigreeFunction apresentaram baixa correlação com o desfecho, demonstrando menor relevância clínica isolada.

3. Abstração

A abstração, no contexto do pensamento computacional, consiste em focar nos elementos mais importantes do problema, deixando de lado informações que não contribuem diretamente para a análise ou para a tomada de decisão. Durante a EDA, a abstração foi aplicada ao selecionar as variáveis mais relevantes para compreender o risco de diabetes, priorizando aquelas com maior impacto clínico e estatístico — como Glicose, IMC e Idade — e deixando em segundo plano atributos de menor poder explicativo, como DiabetesPedigreeFunction e Pressão Diastólica isolada.

4. Algoritmos

O processo foi estruturado como um algoritmo reproduzível em Python, utilizando bibliotecas como pandas, numpy, seaborn e matplotlib. Esse algoritmo representa um **fluxo lógico de análise** — passo a passo — desde o carregamento até a interpretação

inicial dos dados. Cada etapa foi organizada para garantir consistência, clareza e eficiência na análise.

Metadados (Descrição do dataset)

O dataset contém 768 registros de mulheres Pima, divididas entre 268 (35%) diagnosticadas com diabetes e 500 (65%) não diagnosticadas.

Tipo de arquivo:

O conjunto de dados encontra-se disponível em formato CSV, adequado para manipulação em softwares de análise de dados e linguagens de programação como Python e R.

Origem dos dados:

Conjunto de dados aberto (público) do NIDDK (EUA), com informações de mulheres da tribo indígena Pima do Arizona.

Sensibilidade:

Os dados apresentados são sem identificação e não contêm informações que permitam a identificação individual dos participantes. Entretanto, por se tratar de dados clínicos e de saúde, classificam-se como dados sensíveis segundo a Lei Geral de Proteção de Dados Pessoais, ainda que sem identificação.

Validade:

Os dados do Pima Indians Diabetes Dataset não têm validade temporal atual, pois foram coletados há décadas (anos 1970–1980) e refletem uma população específica (mulheres Pima). Ainda assim, o conjunto continua válido para fins educacionais e experimentais, mas não é adequado para uso clínico ou decisões reais de saúde.

Proprietário do dado:

O proprietário e responsável original pelo estudo é o National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), órgão vinculado ao National Institutes of Health (NIH), dos Estados Unidos. O dataset foi posteriormente disponibilizado em plataformas públicas, como o Kaggle e o UCI Machine Learning Repository, para fins acadêmicos e de pesquisa.

Informações adicionais:

Restrições de uso:

De acordo com o repositório Kaggle, o conjunto de dados encontra-se sob licença pública (CC0: Public Domain), o que permite sua utilização, modificação e redistribuição sem necessidade de autorização prévia. No entanto, recomenda-se observar as diretrizes éticas e legais previstas na LGPD, em especial no tratamento e divulgação de dados sensíveis, mesmo quando anonimizados.

Definição de atributos:

São 8 variáveis de entrada (numéricas) e 1 variável alvo:

Pregnancies (int): número de gestações.

Glucose (int): concentração de glicose plasmática.

BloodPressure (int): pressão arterial diastólica (mmHg).

SkinThickness (int): espessura da dobra cutânea tricipital (mm).

Insulin (int): concentração de insulina sérica (mu U/ml).

BMI (float): índice de massa corporal (kg/m²).

DiabetesPedigreeFunction (float): risco calculado com base em histórico familiar.

Age (int): idade em anos.

Outcome (int): variável de saída (0 = não diabético; 1 = diabético).

Observações sobre os dados:

Algumas variáveis contêm valores zero que não são plausíveis (ex.: pressão arterial ou IMC iguais a zero). Isso indica registros incompletos ou dados mascarados, exigindo tratamento antes de análises preditivas.

O dataset apresenta desbalanceamento de classes, o que deve ser considerado no uso de modelos de aprendizado de máquina.

Análise Exploratória dos Dados

O dataset utilizado contém informações clínicas de mulheres indígenas Pima, coletadas com o objetivo de investigar fatores associados ao diabetes tipo 2. As variáveis representam medidas biométricas e laboratoriais de interesse clínico. A seguir, são apresentados os valores de referência clínica para cada atributo, de acordo com diretrizes médicas reconhecidas.

Tabela 1 – Valores de referência clínica das variáveis analisadas

Atributo	Descrição	Faixa de Referência / Interpretação Clínica
Glucose	Concentração de glicose plasmática em jejum (mg/dL)	Normal: < 100 mg/dL Pré-diabetes: 100–125 mg/dL Diabetes: ≥ 126 mg/dL
BloodPressure	Pressão arterial diastólica (mmHg)	Normal: 60–80 mmHg Pré-hipertensão: 81–89 mmHg Hipertensão: ≥ 90 mmHg
SkinThickness	Espessura da dobra cutânea tricipital (mm)	Mulheres: 16–23 mm (normal) Homens: 10–18 mm (normal) > 25 mm → maior adiposidade subcutânea
Insulin	Concentração de insulina sérica em jejum (μU/mL)	Normal: 2–25 μU/mL Resistência insulínica: > 25 μU/mL Hiperinsulinemia acentuada: > 50 μU/mL
BMI (Body Mass Index)	Índice de Massa Corporal (kg/m ²)	Abaixo do peso: < 18,5 Peso normal: 18,5–24,9 Sobrepeso: 25–29,9 Obesidade I: 30–34,9 Obesidade II: 35–39,9 Obesidade III: ≥ 40

Fonte: American Diabetes Association (ADA, 2024); World Health Organization (WHO, 2023); Sociedade Brasileira de Cardiologia (SBC, 2023).

Para realizar a análise exploratória dos dados buscamos desenvolver um script com a linguagem Python para que pudéssemos entender e manipular os dados a fim de obter respostas sobre esses dados. Utilizamos também a ferramenta Google Colab para este projeto.

As bibliotecas que utilizamos para a análise são: pandas, numpy, matplotlib, seaborn, sklearn e scipy.

Imagem 1 – Bibliotecas utilizadas

```
# ===== BIBLIOTECAS =====  
import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sns  
from sklearn.impute import KNNImputer  
from scipy import stats
```

Inicialmente na análise queremos entender a estrutura dos dados, que se encontram em formato tabular. Para isso desenvolvemos o seguinte trecho para apresentar as cinco primeiras linhas:

Imagem 2 – Script para leitura e apresentação da estrutura dos dados

```
# ===== LEITURA DOS DADOS =====  
df = pd.read_csv('pima_dataset.csv')  
  
print("Dimensões do dataset:", df.shape)  
print("\nVisualização inicial:")  
print(df.head())
```

Dessa forma, resultou-se na seguinte estrutura dos dados:

Imagem 3 – Estrutura dos dados

Dimensões do dataset: (768, 9)

Visualização inicial:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	\
0	6	148	72	35	0	33.6	
1	1	85	66	29	0	26.6	
2	8	183	64	0	0	23.3	
3	1	89	66	23	94	28.1	
4	0	137	40	35	168	43.1	

	DiabetesPedigreeFunction	Age	Outcome
0	0.627	50	1
1	0.351	31	0
2	0.672	32	1
3	0.167	21	0
4	2.288	33	1

Temos ao todo 768 registros (ou seja, 768 mulheres Pima) registradas no dataset, contendo 9 colunas, sendo 8 atributos e 1 variável alvo (outcome).

Para cada atributo notou-se a necessidade de entender o tipo de dado, portanto para este fim incluímos outro trecho informativo.

Imagem 4 – Script de informações básicas

```
# ===== INFORMAÇÕES BÁSICAS =====
print("\nInformações gerais:")
print(df.info())

print("\nResumo estatístico:")
print(df.describe())
```

Imagem 5 – Informações gerais (tipo e contagem de nulos)

```

Informações gerais:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Pregnancies            768 non-null    int64
1   Glucose                768 non-null    int64
2   BloodPressure          768 non-null    int64
3   SkinThickness          768 non-null    int64
4   Insulin                768 non-null    int64
5   BMI                    768 non-null    float64
6   DiabetesPedigreeFunction 768 non-null    float64
7   Age                    768 non-null    int64
8   Outcome                768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
None

```

Observa-se que o registro inicia-se no 0 e segue até 767, sendo cada registro uma mulher Pima conforme já citado anteriormente.

Novamente temos a informação do número de colunas, sendo 9 ao todo. Apenas BMI e DiabetesPedigreeFunction são do tipo float, sendo as demais do tipo int.

Dentro do script de informações básicas também verificamos um resumo estatístico dos dados, contendo medidas de dispersão, para entender melhor os valores que estão contidos no dataset. Temos informações como contagem, média, desvio padrão e quartis.

Imagem 6 – Resumo estatístico

Resumo estatístico:					
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin \
count	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479
std	3.369578	31.972618	19.355807	15.952218	115.244002
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000
75%	6.000000	140.250000	80.000000	32.000000	127.250000
max	17.000000	199.000000	122.000000	99.000000	846.000000

	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000
mean	31.992578	0.471876	33.240885	0.348958
std	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.078000	21.000000	0.000000
25%	27.300000	0.243750	24.000000	0.000000
50%	32.000000	0.372500	29.000000	0.000000
75%	36.600000	0.626250	41.000000	1.000000
max	67.100000	2.420000	81.000000	1.000000

Foi necessário tratar os dados pois haviam dados incorretos, como colunas de glicose, pressão do sangue, insulina, dobra cutânea e IMC, que sem tratamento prejudicam nossa análise.

Imagem 7 – Script de ajuste de dados incorretos

```
# ===== AJUSTE DE VALORES INCORRETOS =====
# Algumas colunas não devem ter valor 0 (ex: pressão, glicose, IMC)
cols_invalidas = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']
for col in cols_invalidas:
    df[col] = df[col].replace(0, np.nan)

print("\nValores ausentes após substituição dos zeros incorretos:")
print(df.isna().sum())
```

Observamos aqui uma grande quantidade de valores faltantes na variável de insulina e dobra cutânea, o que certamente pode

enviesar a análise ou comprometer algoritmos de análise preditiva com resultados distorcidos.

Imagem 8 – Valores ausentes

```
Valores ausentes após substituição dos zeros incorretos:  
Pregnancies          0  
Glucose              5  
BloodPressure        35  
SkinThickness        227  
Insulin              374  
BMI                  11  
DiabetesPedigreeFunction  0  
Age                  0  
Outcome              0  
dtype: int64
```

Após esse processo, realizamos a imputação de valores com base nas características das variáveis, utilizando estratégias como moda, média, mediana e KNN.

Imagem 9 – Script de imputação de dados

```
# ===== TRATAMENTO DE VALORES FALTANTES =====  
# Estratégia mista: média, mediana, moda, e KNN  
  
# 1. Substituir colunas com poucos nulos pela mediana  
for col in ['Glucose', 'BloodPressure', 'BMI']:  
    df[col].fillna(df[col].median(), inplace=True)  
  
# 2. Substituir colunas mais críticas pelo KNN  
imputer = KNNImputer(n_neighbors=5)  
df[['SkinThickness', 'Insulin']] = imputer.fit_transform(df[['SkinThickness', 'Insulin']])  
  
print("\nValores ausentes após imputação:")  
print(df.isna().sum())
```

Com isso, nenhuma variável possui valor faltante.

Imagem 10 – Valores ausentes após imputação

```
Valores ausentes após imputação:  
Pregnancies      0  
Glucose           0  
BloodPressure     0  
SkinThickness     0  
Insulin           0  
BMI               0  
DiabetesPedigreeFunction  0  
Age               0  
Outcome           0  
dtype: int64
```

Notamos que há alguns valores que fogem do padrão, conhecidos como outliers. Esses valores podem comprometer estimativas que utilizam média, então nesse caso os retiramos.

Imagem 11 – Script de detecção e remoção de outliers

```
# ===== DETECÇÃO DE OUTLIERS =====  
# Usando o método do Z-score  
z_scores = np.abs(stats.zscore(df.select_dtypes(include=[np.number])))  
df_sem_outliers = df[(z_scores < 3).all(axis=1)]  
print(f"\nRemovidos {df.shape[0] - df_sem_outliers.shape[0]} outliers.")  
  
df = df_sem_outliers
```

Foram detectados 48 outliers, dessa forma, foram removidos.

Imagem 12 – Outliers removidos

```
Removidos 48 outliers.
```

Após realizarmos o entendimento da estrutura dos dados e fazermos os devidos tratamentos, queremos entender quantas mulheres possuem diabetes e, para entendermos isso, vejamos a contagem e a representação gráfica.

Imagem 13 – Análise Exploratória

```
# ===== ANÁLISE EXPLORATÓRIA =====
print("\nDistribuição da variável alvo (Outcome):")
print(df['Outcome'].value_counts())

plt.figure(figsize=(5,4))
sns.countplot(data=df, x='Outcome', palette='Set2')
plt.title('Distribuição de Casos de Diabetes (0 = Não, 1 = Sim)')
plt.show()

# Histograma e boxplots
df.hist(bins=20, figsize=(14,10), color='teal')
plt.suptitle('Distribuição das Variáveis Numéricas')
plt.show()

# Boxplots por diabetes
plt.figure(figsize=(14,10))
for i, col in enumerate(df.columns[:-1]):
    plt.subplot(3,3,i+1)
    sns.boxplot(x='Outcome', y=col, data=df, palette='Set2')
    plt.title(col)
plt.tight_layout()
plt.show()
```

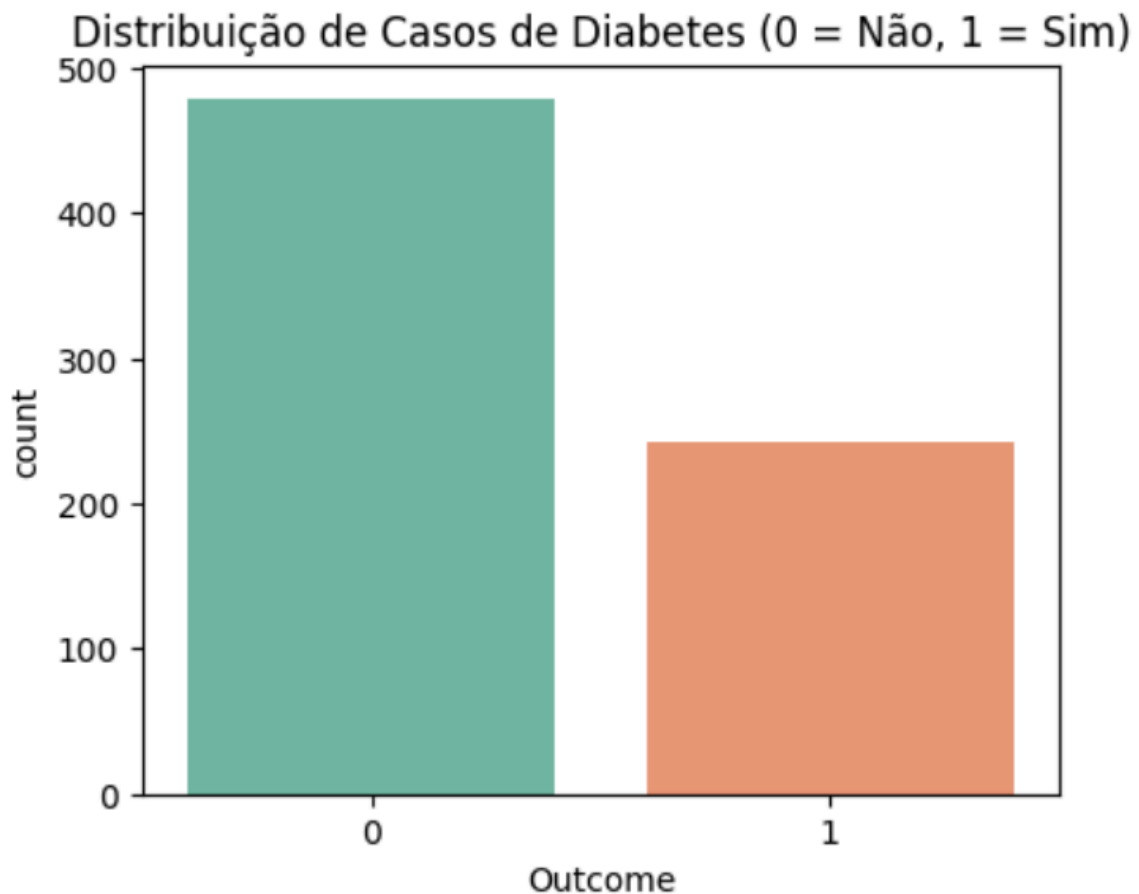
A distribuição se deu por 478 mulheres sem diabetes e 242 mulheres com diabetes. A saída 0 representa uma mulher sem diabetes, 1 representa diabética.

Imagem 14 – Total de casos diabéticos e não diabéticos

```
Distribuição da variável alvo (Outcome):  
Outcome  
0      478  
1      242
```

Graficamente temos uma noção do quanto mulheres diabéticas representam do todo.

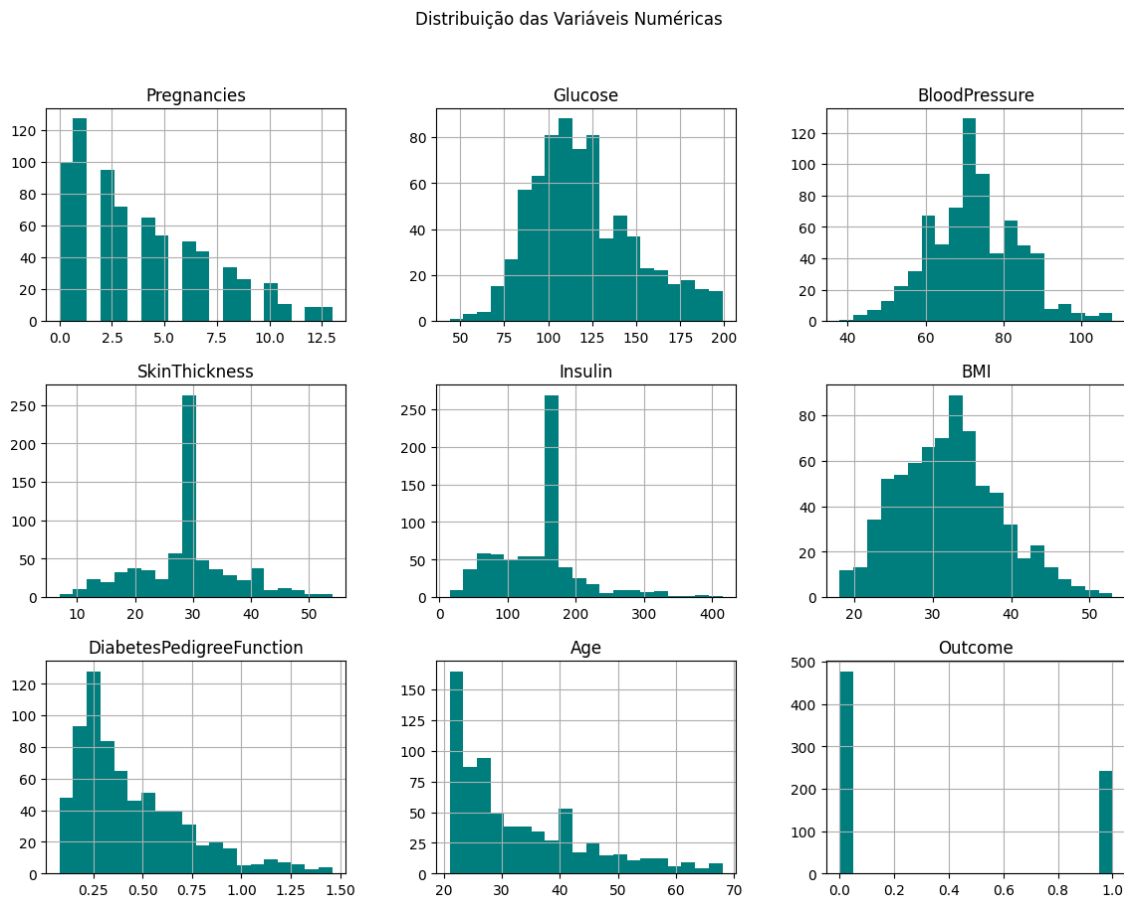
Imagem 15 – Distribuição de casos diabéticos e não diabéticos



O gráfico acima mostra a distribuição dos casos de diabetes (Outcome). Nota-se que há mais pessoas sem diabetes do que com diabetes, o que indica um desbalanceamento nos dados. Esse

detalhe é importante, pois pode influenciar as análises e os modelos preditivos, já que há mais exemplos de uma classe do que da outra.

Imagem 16 – Distribuição das Variáveis Numéricas



O conjunto de gráficos mostra a **distribuição das variáveis numéricas** do conjunto de dados. É possível observar que algumas variáveis, como **Glucose** e **BMI**, seguem uma distribuição mais concentrada em torno de certos valores, indicando uma tendência na amostra. Outras, como **Insulin** e **SkinThickness**, apresentam grande variação e alguns valores extremos (outliers). A variável **Pregnancies** mostra uma queda gradual na frequência, ou seja, poucas pessoas tiveram muitas gestações. Esses gráficos ajudam a

entender o comportamento geral dos dados e identificar possíveis valores fora do padrão.

Dessa forma, a maioria das variáveis apresenta distribuições assimétricas, com concentração de valores em faixas específicas.

A variável Glucose (nível de glicose no sangue) destaca-se como a mais relevante, apresentando valores mais elevados em indivíduos com diagnóstico positivo para diabetes. BMI (índice de massa corporal) e Age (idade) também mostram relação significativa, indicando que o sobrepeso e o avanço da idade aumentam o risco da doença.

As variáveis Pregnancies (número de gestações) e DiabetesPedigreeFunction (histórico familiar) reforçam a influência de fatores reprodutivos e genéticos. Já BloodPressure, SkinThickness e Insulin apresentam distribuições mais dispersas ou concentradas, sugerindo menor impacto direto, embora ainda relevantes para o contexto clínico.

Por fim, a variável Outcome evidencia um desequilíbrio de classes, com maior número de pessoas não diabéticas, o que deve ser considerado na modelagem preditiva.

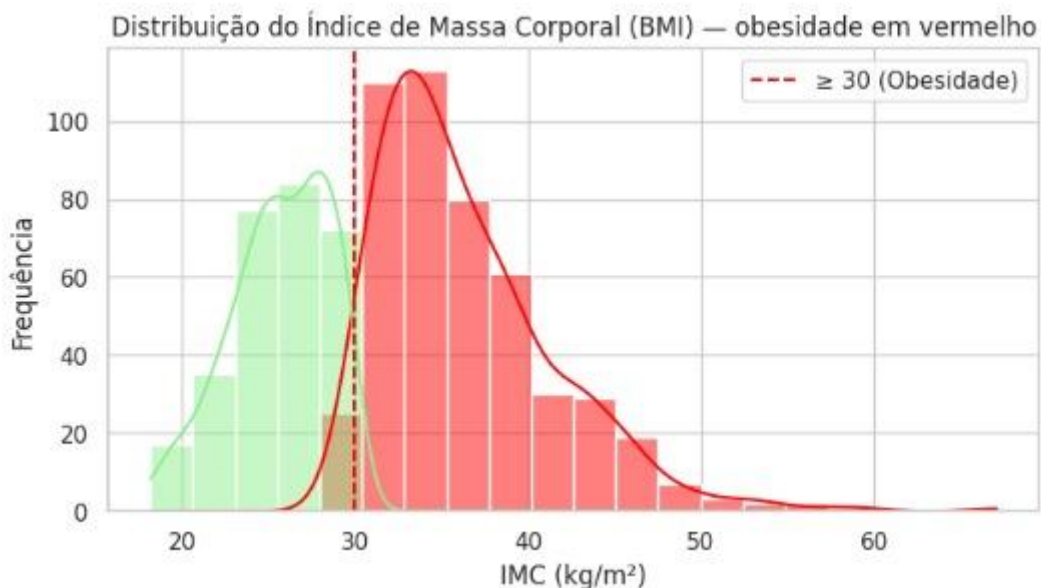
Em síntese, os gráficos indicam que níveis elevados de glicose, IMC, idade e número de gestações estão entre os principais fatores associados à presença de diabetes no conjunto analisado.

Analisemos agora o gráfico abaixo que apresenta a distribuição do IMC das pacientes, destacando em vermelho os valores iguais ou superiores a **30 kg/m²**, classificados como obesidade segundo a

OMS. Observa-se que uma parte significativa das mulheres ultrapassa esse limite, mostrando forte prevalência de sobrepeso e obesidade na população analisada.

Como o IMC é um dos marcadores metabólicos relacionados ao risco de diabetes tipo 2, essa concentração de valores elevados reforça o padrão identificado em outras análises do estudo.

Imagem 17 – Distribuição do Índice de Massa Corporal (IMC)

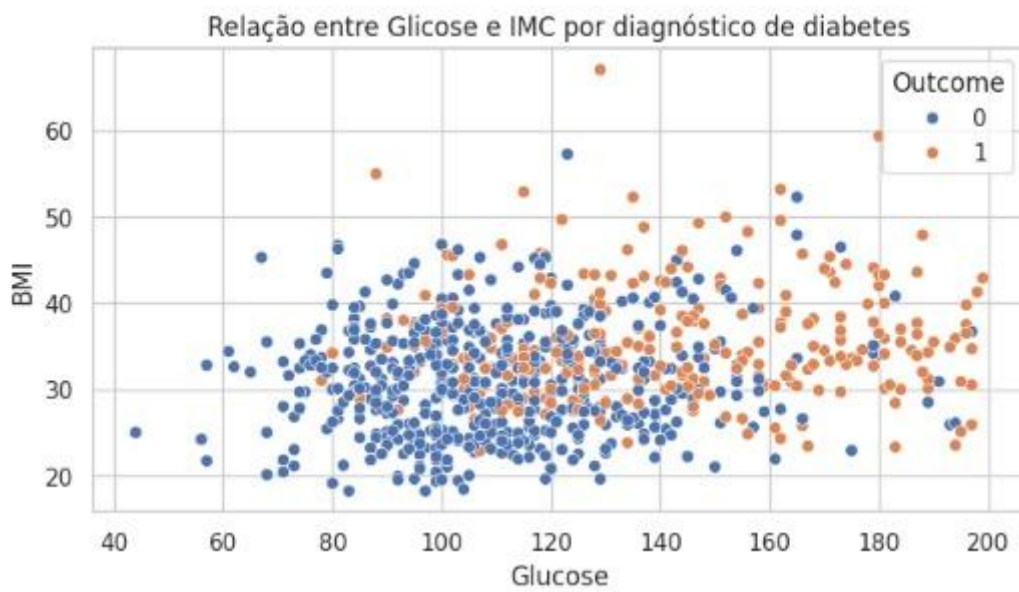


No gráfico de dispersão a seguir que apresenta a relação entre glicose e IMC, nota-se que as pacientes diagnosticadas com diabetes (**Outcome = 1**) apresentam maior concentração de pontos na região de **valores elevados de glicose e IMC acima da média**.

Esse comportamento indica um padrão conjunto: **quanto maior a glicose e o IMC, maior tende a ser a probabilidade de ocorrência de diabetes**.

A presença de pacientes não diabéticos com valores mais baixos reforça a separação visual entre os grupos.

Imagem 18 – Gráfico de dispersão da relação entre Glicose e IMC por diagnóstico de diabetes

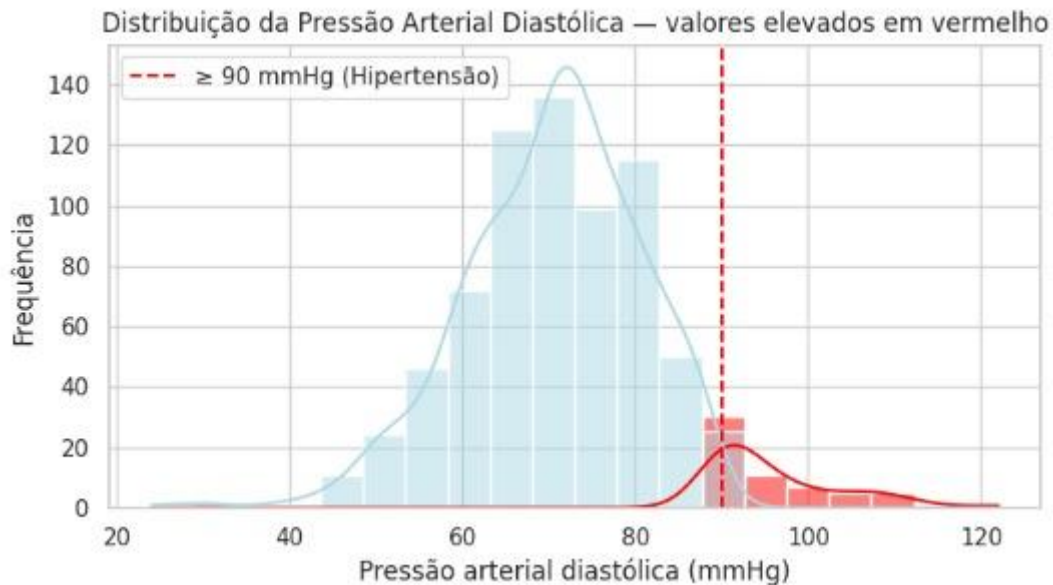


Ao tentar entender mais sobre a influência da hipertensão em casos positivos de diabetes geramos o gráfico que mostra a distribuição dos valores de pressão arterial diastólica, destacando em vermelho as medições ≥ 90 mmHg, que são tradicionalmente classificadas como hipertensão.

Embora a maior parte da população esteja dentro da faixa normal, existe um grupo de pacientes que ultrapassa esse limite, sugerindo um risco adicional relacionado à saúde cardiovascular, que frequentemente se associa ao diabetes e a síndromes metabólicas.

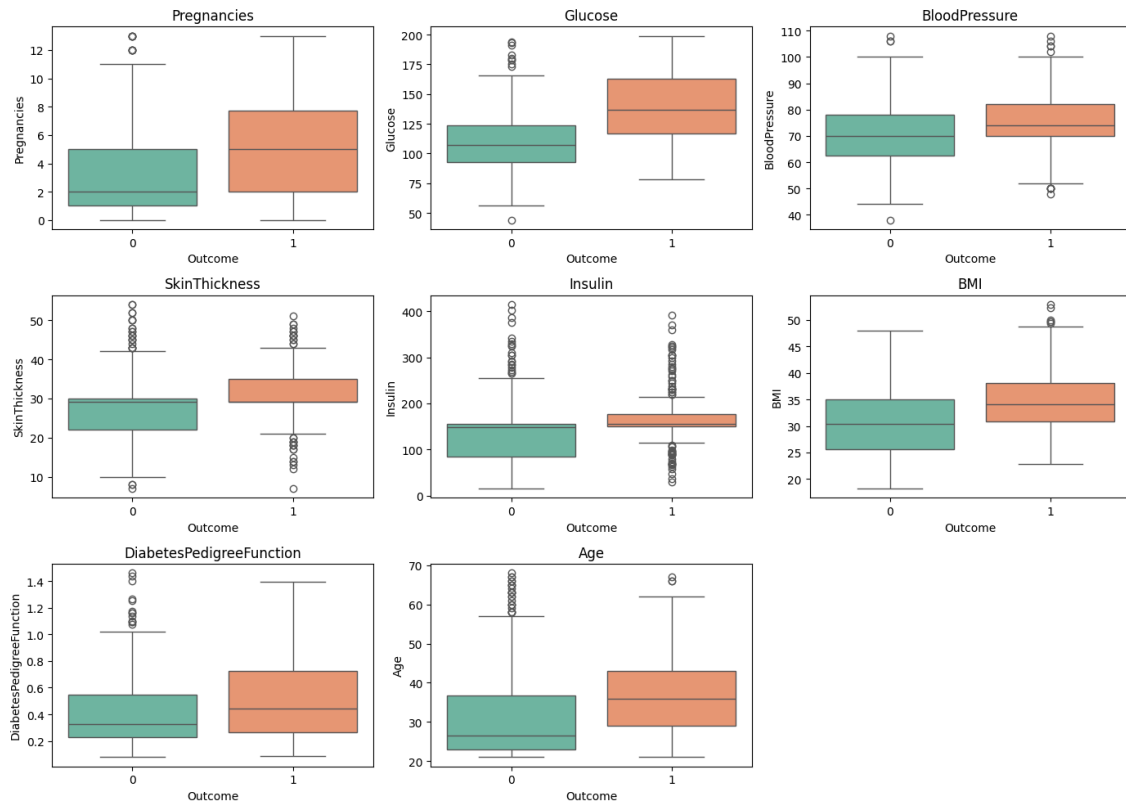
Esse resultado indica que, apesar de não ser o fator mais determinante isoladamente, a pressão arterial elevada deve ser considerada como parte do contexto clínico dessas mulheres.

Imagem 19 – Distribuição da Pressão Arterial Diastólica



Agora abordando gráficos de boxplots, que permitem visualizar as diferenças estatísticas entre os grupos com e sem diabetes. De modo geral, observa-se que os indivíduos diagnosticados com diabetes (Outcome = 1) apresentam valores medianos mais altos nas variáveis Glucose, BMI, Age e Pregnancies, sugerindo forte relação desses fatores com o desenvolvimento da doença.

Imagem 20 – Boxplots das Variáveis Numéricas



A variável Glucose apresenta a diferença mais expressiva, com valores significativamente maiores entre os diabéticos, reforçando sua importância como indicador principal. O BMI (índice de massa corporal) também é mais elevado nesse grupo, indicando a influência do excesso de peso. Da mesma forma, Age e Pregnancies mostram tendência de aumento entre os casos positivos, o que sugere que a idade avançada e o número de gestações podem contribuir para o risco de diabetes.

Já as variáveis BloodPressure, SkinThickness, Insulin e DiabetesPedigreeFunction apresentam distribuições mais semelhantes entre os grupos, embora ainda revelem ligeiras elevações entre os diabéticos. Além disso, é possível notar a

presença de outliers (pontos fora das caixas) em várias variáveis, o que indica variações individuais relevantes e reforça a importância de tratamento e normalização dos dados antes da modelagem.

Em síntese, os boxplots evidenciam que níveis elevados de glicose e IMC estão diretamente associados à ocorrência de diabetes, enquanto as demais variáveis exercem influência mais moderada.

Dando sequência a análise exploratória, queremos entender também a correlação entre as variáveis, principalmente entre a variável alvo e as demais já citadas, em destaque para a correlação entre diabetes (outcome) com glicose (glucose) e IMC (BMI).

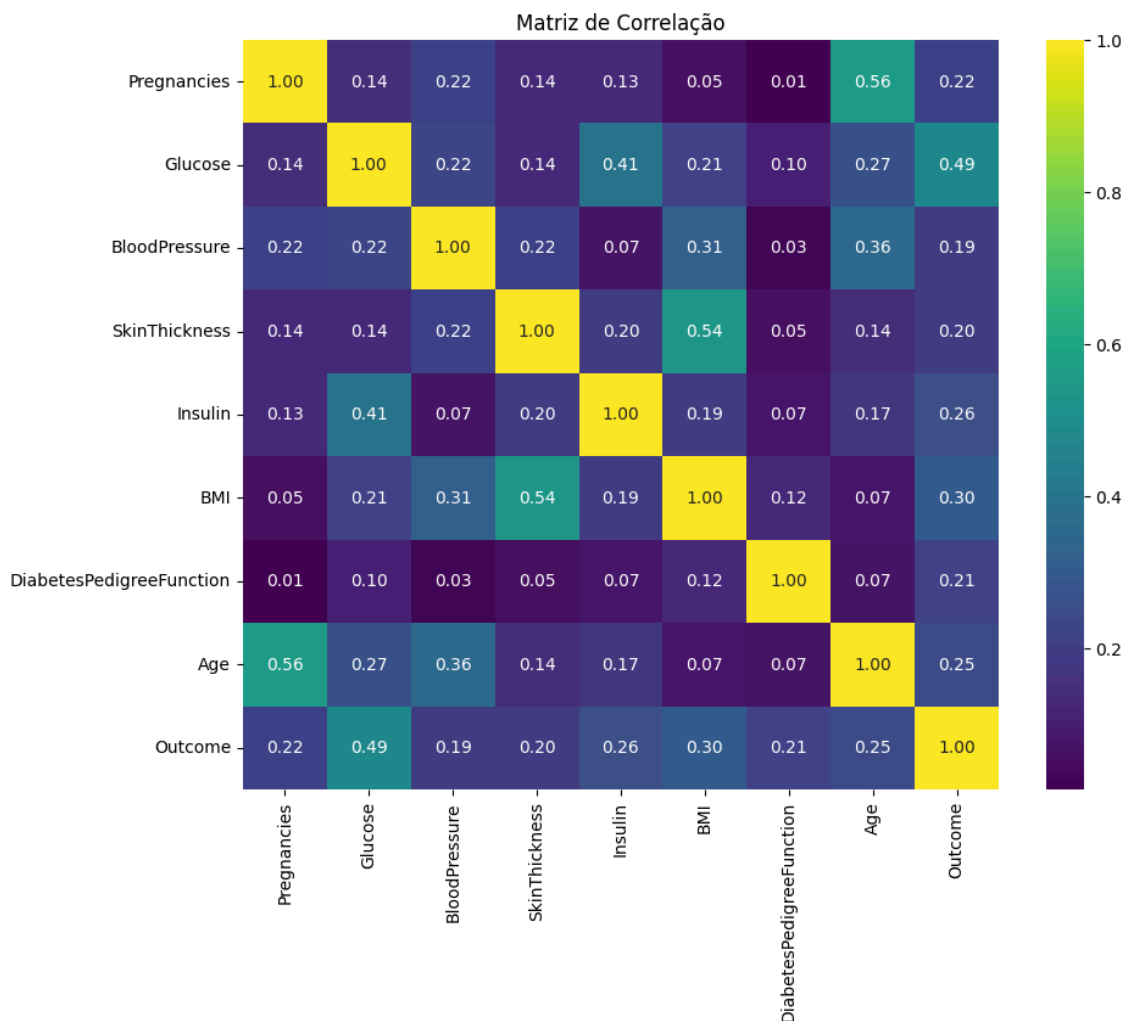
O script abaixo executa a geração de uma matriz de correlação:

Imagem 21 – Script de geração da matriz de correlação

```
# ===== CORRELAÇÕES =====  
plt.figure(figsize=(10,8))  
sns.heatmap(df.corr(), annot=True, cmap='viridis', fmt=".2f")  
plt.title('Matriz de Correlação')  
plt.show()
```

Através da matriz que se segue abaixo podemos observar as correlações entre as variáveis.

Imagem 22 – Matriz de correlação entre variáveis



A matriz de correlação mostra o grau de relação entre as variáveis do conjunto de dados. Observa-se que a variável Glucose (0.49) apresenta a correlação mais forte com o Outcome, indicando que níveis mais altos de glicose estão diretamente associados ao diagnóstico de diabetes.

Outras variáveis com correlação moderada com o Outcome são BMI (0.30) e Age (0.25), sugerindo que o aumento do índice de massa corporal e da idade também eleva o risco da doença.

Além disso, há correlação relevante entre SkinThickness e BMI (0.54), o que faz sentido, já que ambos estão ligados à composição corporal.

De forma geral, a matriz reforça que fatores como glicose e IMC são os mais influentes para prever o diabetes neste conjunto de dados.

Além da matriz optamos por resumir a correlação do diabetes com as demais variáveis com o objetivo de extrair insights para apoiar decisões que possam ser necessárias em casos clínicos como o abordado neste projeto.

Imagem 23 – Script de resumo de correlação

```
# ===== INSIGHTS =====  
corr_target = df.corr()['Outcome'].sort_values(ascending=False)  
print("\nCorrelação das variáveis com o diabetes:")  
print(corr_target)
```

Imagem 24 – Resumo de correlação das variáveis

```
Correlação das variáveis com o diabetes:  
Outcome          1.000000  
Glucose           0.487006  
BMI               0.303676  
Insulin           0.255045  
Age               0.246310  
Pregnancies       0.218373  
DiabetesPedigreeFunction 0.209538  
SkinThickness     0.202164  
BloodPressure     0.189198  
Name: Outcome, dtype: float64
```

O resumo acima confirma o que extraímos na matriz de correlação, só que trazendo dados mais precisos, com mais casas decimais, diferente da matriz que arredonda os valores, destacando a variável Glucose com 0.487 e BMI em seguida com 0.303.

Outra forma para confirmar a relação da diabetes com as variáveis foi com a utilização de pairplot para entender o seu comportamento.

Imagem 25 – Script de criação de pairplot

```
# ===== RELAÇÕES ESPECÍFICAS =====  
sns.pairplot(df, hue='Outcome', diag_kind='kde', corner=True, palette='husl')  
plt.suptitle('Relação entre Variáveis e Diagnóstico de Diabetes', y=1.02)  
plt.show()  
  
print("\nAnálise concluída com sucesso!")
```

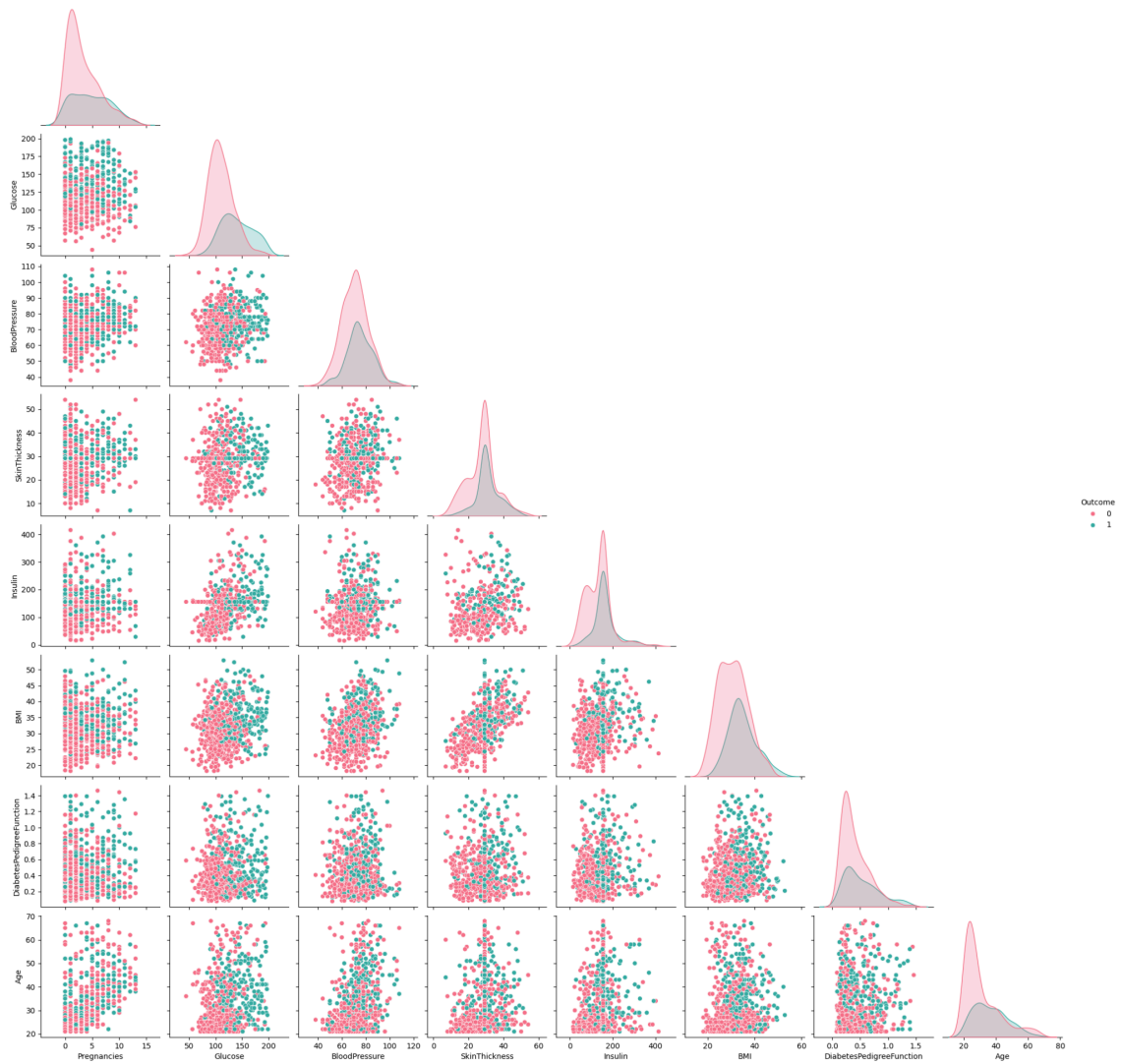
O pairplot mostra como as variáveis se relacionam entre si e com o diagnóstico de diabetes. É possível observar que pessoas diagnosticadas com diabetes (Outcome = 1, em rosa) tendem a apresentar valores mais altos de glicose e IMC, o que reforça o padrão visto na matriz de correlação.

As distribuições também indicam que há grupos distintos de pacientes nessas variáveis, enquanto outras, como pressão arterial e espessura da pele, apresentam sobreposição entre os grupos, sugerindo menor impacto direto.

No geral, o gráfico evidencia que o aumento da glicose, do IMC e da idade são fatores mais associados à presença de diabetes, enquanto as demais variáveis mostram relações mais fracas ou indiretas.

Imagem 26 – Pairplot

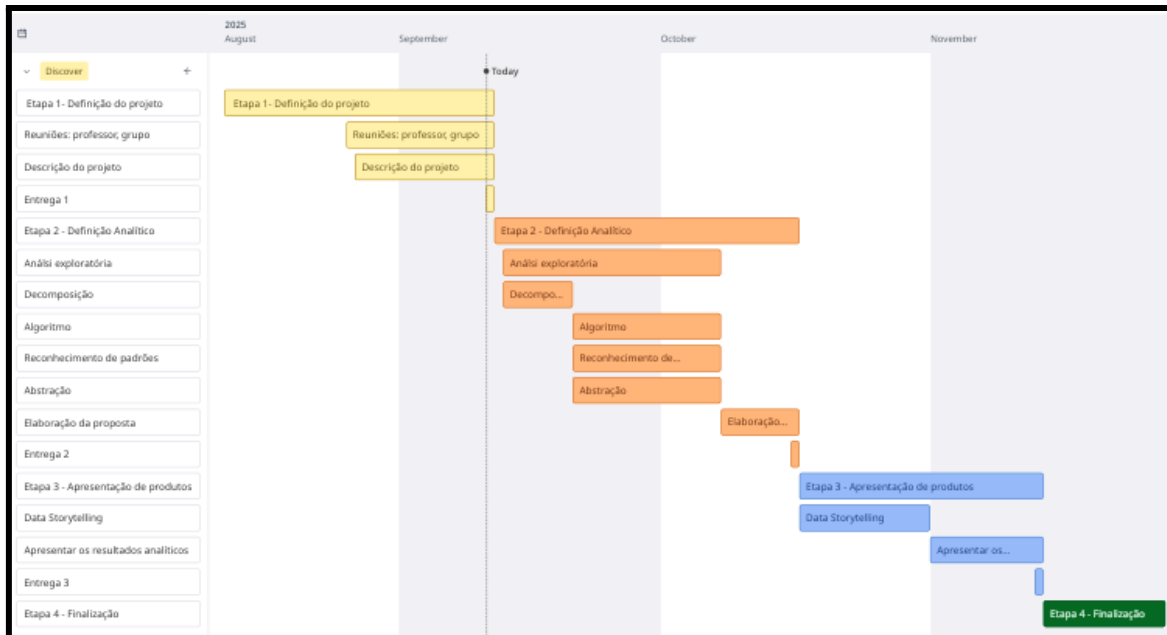
Relação entre Variáveis e Diagnóstico de Diabetes



Cronograma

Para o projeto, o primeiro modelo de cronograma está abaixo:

Imagem 27 – Cronograma de execução



Referências Bibliográficas

NATIONAL INSTITUTE OF DIABETES AND DIGESTIVE AND KIDNEY DISEASES. KAGGLE. *Pima Indians Diabetes Dataset*. Disponível em: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>. Acesso em: 8 set. 2025.

AOBESIDADE. Obesidade mórbida: o paradoxo de Pima. Disponível em: <https://aobesidade.com.br/obesidade-morbida/#:~:text=O%20paradoxo%20de%20Pima&text=Mas%20veja%20a%20seguir%20um,os%20%C3%ADndios%20Pima%20do%20Arizona>. Acesso em: 9 set. 2025.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. Diabetes in the Pima Indians. In: *The Genetic Basis of Common Diseases*. Bethesda: NCBI Bookshelf, 1992. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK233089/>. Acesso em: 9 set. 2025.

NATIONAL INSTITUTES OF HEALTH. NIH's work with Native communities drives diabetes research. In: *NIH Intramural Research Program Catalyst*, v. 29, n. 6, 2021. Disponível em: <https://irp.nih.gov/catalyst/29/6/nihs-work-with-native-communities-drives-diabetes-research>. Acesso em: 9 set. 2025.

PUBMED. KNOWLER, W. C.; BENNETT, P. H.; HANSON, R. L.; et al. Diabetes incidence and prevalence in Pima Indians: influence of obesity and family history. *Diabetes Care*, v. 16, n. 1, p. 120–126,

1993. DOI: 10.2337/diacare.16.1.120. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/8422779/>. Acesso em: 9 set. 2025.

PUBMED. ESCOBEDO, J. et al. Prevalence of diabetes and obesity in the Pima Indians of Mexico. *Diabetes Care*, v. 29, n. 8, p. 1852–1856, 2006. DOI: 10.2337/dc06-0040. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/16873794/>. Acesso em: 9 set. 2025.

RESEARCHGATE. Pima Indian Diabetes dataset attributes. In: *Article: A comparative study of classification algorithms for diabetes prediction*. Disponível em: https://www.researchgate.net/figure/Pima-Indian-Diabetes-dataset-attributes_tbl1_325653625. Acesso em: 9 set. 2025.

SCIENTIFIC AMERICAN / SAGE. Pima Indians and Obesity. In: *Encyclopedia of Obesity*. Thousand Oaks: Sage Publications, 2008. Disponível em: <https://sk.sagepub.com/ency/edvol/embed/obesity/chpt/pima-indians>. Acesso em: 9 set. 2025.

SCIENTIFIC DIRECT. KUMAR, A. et al. An interpretable machine learning model for diabetes prediction using the Pima Indians dataset. *Heliyon*, v. 10, n. 1, 2024. DOI: 10.1016/j.heliyon.2024.e23957. Disponível em: <https://www.sciencedirect.com/science/article/pii/S240584402400567X>. Acesso em: 9 set. 2025.

NATIONAL INSTITUTE OF DIABETES AND DIGESTIVE AND KIDNEY DISEASES. *Pima Indian Diabetes*. In: *The Genetic Basis of*

Common Diseases. Bethesda: NCBI Bookshelf, 1992. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK233089/>. Acesso em: 8 out. 2025.

NATIONAL INSTITUTES OF HEALTH. *NIH's work with Native communities drives diabetes research*. In: NIH Intramural Research Program Catalyst, v. 29, n. 6, 2021. Disponível em: <https://irp.nih.gov/catalyst/29/6/nihs-work-with-native-communities-drives-diabetes-research>. Acesso em: 9 set. 2025.

UNIVERSITY OF CALIFORNIA IRVINE (UCI). *Pima Indians Diabetes Database*. Disponível em: <https://archive.ics.uci.edu/dataset/34/diabetes>. Acesso em: 8 out. 2025.

Glossário

Diabetes tipo 2: doença crônica que se desenvolve ao longo do tempo, associada a fatores como hábitos alimentares inadequados, sedentarismo e obesidade. O organismo produz insulina, mas há resistência a ela. No diabetes tipo 1, o indivíduo já nasce com predisposição e o corpo não produz insulina.

Doenças crônicas: condições que duram muito tempo, geralmente sem cura definitiva, exigindo tratamento contínuo (ex.: hipertensão, diabetes, doenças cardíacas).

Fatores clínicos: características do paciente importantes para avaliação de risco ou diagnóstico. Exemplo: glicose elevada, pressão arterial alta, colesterol alterado, sobrepeso, histórico familiar.

Glicose: açúcar que serve de principal fonte de energia para o corpo. Quando não é bem regulada (por falha na ação da insulina), pode aumentar no sangue, caracterizando risco ou presença de diabetes.

Índice de Massa Corporal (BMI): medida para avaliar se o peso de uma pessoa está na faixa saudável. Calculado por massa/peso (kg) dividido pelo quadrado da altura (m^2). Altos valores indicam sobrepeso ou obesidade.

Insulina: hormônio produzido pelo pâncreas, que permite que a glicose entre nas células e seja usada como energia. Em DM2, seu efeito é prejudicado; em DM1, ela não é produzida.

Morbidade: probabilidade de desenvolvimento de outras doenças ou complicações associadas a uma condição de base (como a diabetes), por exemplo problemas cardíacos, renais, visuais ou circulatórios.

Mortalidade: capacidade ou probabilidade de uma doença levar à morte — no caso da diabetes, pelas complicações graves como infarto, acidente vascular cerebral, insuficiência renal ou infecções.

Pressão arterial: força que o sangue exerce nas paredes das artérias. A pressão alta é comum em pessoas com diabetes e aumenta o risco de complicações cardiovasculares.

Prevalência: proporção de pessoas que têm uma determinada doença em uma população num dado momento.

Saúde metabólica: equilíbrio de funções metabólicas do organismo — glicose, lipídios, pressão arterial, gordura corporal etc. — de modo que reduz o risco de doenças crônicas.

Vulneráveis: indivíduos ou grupos com maior risco de adoecer ou de sofrer complicações, por razões sociais, biológicas, econômicas ou de acesso aos cuidados de saúde.