📄 Documentação do Projeto: Algoritmo de Machine Learning - Regressão Logística para Diagnóstico de Diabetes

Sumário

---

1. Introdução, Desafio e Objetivos

Um grande hospital universitário busca implementar um sistema inteligente de suporte ao diagnóstico, capaz de ajudar médicos e equipes clínicas na análise inicial de exames e no processamento de dados médicos.

O presente trabalho visa o desenvolvimento e a validação de do modelo de algoritmo preditivo Regressão Logística, que baseado em Machine Learning para auxiliar no diagnóstico de diabetes. O objetivo principal é criar um algoritmo capaz de classificar exames médicos e documentos clínicos de pacientes com alta **eficácia**.

2. Aquisição e Descrição do Dataset

O projeto utilizou o **Dataset sobre diabetes**.

- **Fonte:** N. Inst. of Diabetes & Diges. & Kidney Dis.
- **Amostras:** O dataset é composto por **768** observações de pacientes.

Pima Indians Diabetes

Contexto

Este conjunto de dados é originário do **Instituto Nacional de Diabetes e Doenças Digestivas e Renais** (National Institute of Diabetes and Digestive and Kidney Diseases). O objetivo é prever, com base em medições de diagnóstico, se uma paciente tem diabetes.

Conteúdo

Em particular, todas as pacientes aqui são do sexo **feminino**, com pelo menos **21 anos** de idade e de **ascendência indígena Pima**.

Mais informações sobre a tribo.

https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data

- Abaixo as 5 mineiras linhas do dataset

```
Primeiras 5 linhas do dataset:
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  DiabetesPedigreeFunction  Age  Outcome
0            6      148             72             35        0  33.6                     0.627   50        1
1            1       85             66             29        0  26.6                     0.351   31        0
2            8      183             64              0        0  23.3                     0.672   32        1
3            1       89             66             23       94  28.1                     0.167   21        0
4            0      137             40             35      168  43.1                     2.288   33        1
```

- **Características (Features):**

```
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
```

- **Variável Alvo:** A variável de interesse é o diagnóstico **"Outcome"** (0 para Não-Diabético, 1 para Diabético).

- **Estatísticas descritivas:**

```
\Estatísticas descritivas (incluindo Mínimo e Máximo):
       Pregnancies     Glucose  BloodPressure  ...  DiabetesPedigreeFunction         Age     Outcome
count   768.000000  768.000000     768.000000  ...                768.000000  768.000000  768.000000
mean      3.845052  120.894531      69.105469  ...                  0.471876   33.240885    0.348958
std       3.369578   31.972618      19.355807  ...                  0.331329   11.760232    0.476951
min       0.000000    0.000000       0.000000  ...                  0.078000   21.000000    0.000000
25%       1.000000   99.000000      62.000000  ...                  0.243750   24.000000    0.000000
50%       3.000000  117.000000      72.000000  ...                  0.372500   29.000000    0.000000
75%       6.000000  140.250000      80.000000  ...                  0.626250   41.000000    1.000000
max      17.000000  199.000000     122.000000  ...                  2.420000   81.000000    1.000000
```

3. Análise Exploratória de Dados (EDA)

A análise exploratória revelou a distribuição das variáveis e a presença de desafios críticos que necessitaram de tratamento:

- **Problema 1 - Contagem de Nulos (NaN):** Observou-se a presença de valores zero (0) em colunas que não deveriam aceitar tal valor:

  Os totais verificados NaN por feature:
  Destaque para a maior quantidade de nulos para **SkinThickness** e **Insulin**

  Insight 1

  SkinThickness (espessura da pele)

  A relação entre a espessura da pele e o diagnóstico de diabetes é complexa, pois o diabetes pode causar **diferentes alterações na pele**, que incluem tanto o **espessamento** quanto o **afinamento/ressecamento** em áreas específicas.

Insight 2

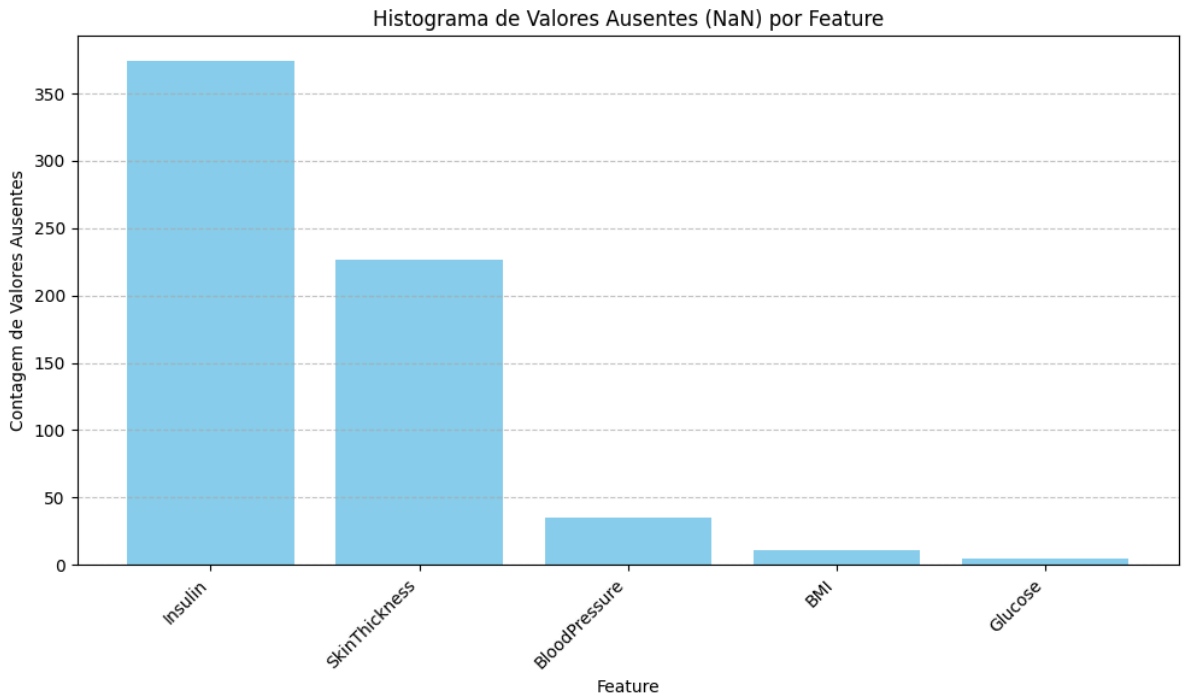
Insulin (insulina)

A relação da **quantidade de insulina** com o diagnóstico de diabetes é fundamental, pois o diabetes mellitus (DM) é caracterizado por uma produção insuficiente ou pela má utilização da insulina pelo organismo.

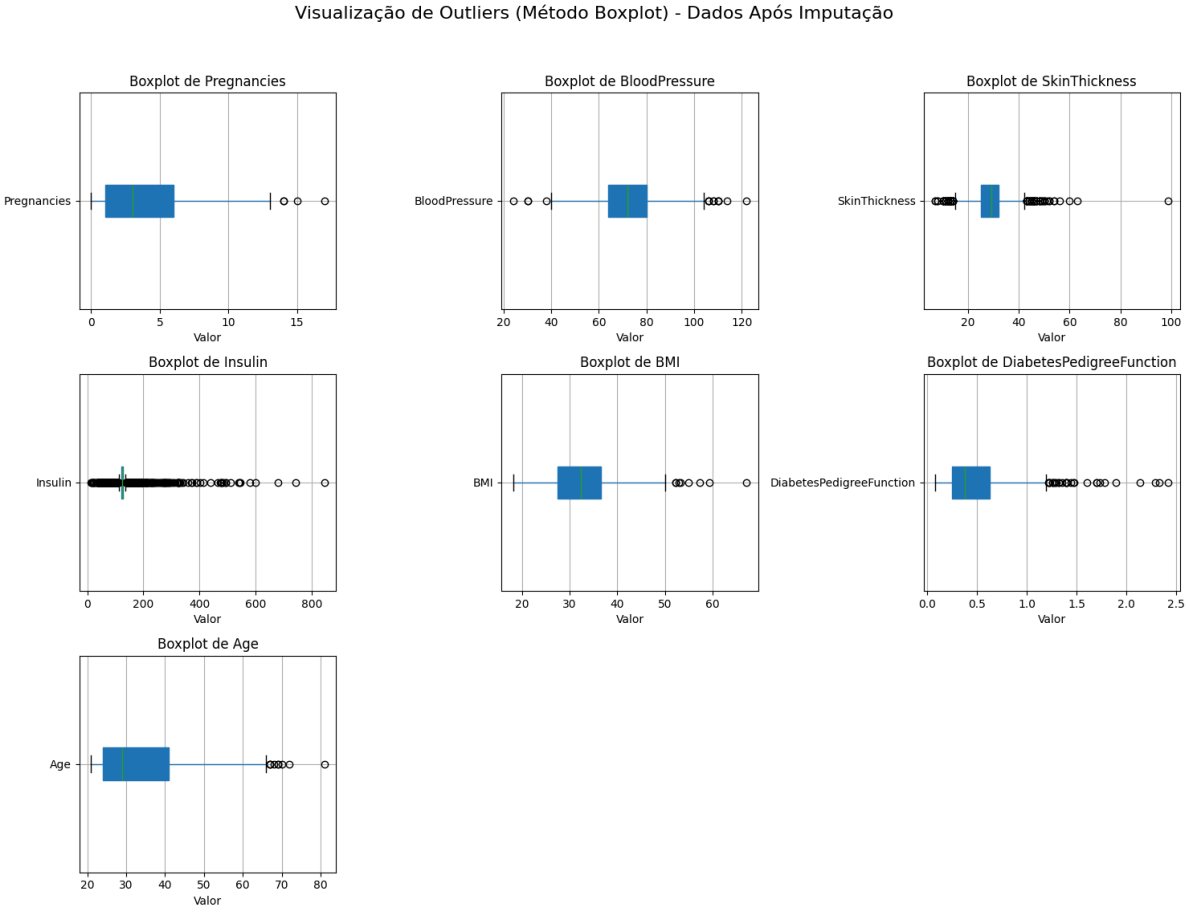
A forma como essa relação se manifesta difere entre os principais tipos de diabetes: Tipo 1 (DM1) e Tipo 2 (DM2)

```
Contagem de valores nulos (NaN) após substituição de zeros:
Pregnancies                 0
Glucose                     5
BloodPressure              35
SkinThickness             227
Insulin                   374
BMI                        11
```

- *Evidência complementar:* Gráficos de histogramas evidenciando os *zeros* anômalos.



- **Problema 2 - Detecção de Outliers:** Usando IQR foram detectados outliers nas colunas: ['Pregnancies', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age'].



- **Problema 3 - Desbalanceamento:**
    - A análise das estatísticas descritivas revela que as diferentes características do dataset estão em escalas variadas e exibem diferentes níveis de dispersão.
    - O Glucose e o Insulin têm valores em torno de 100, enquanto DiabetesPedigreeFunction varia entre 0 e 1.
    - As colunas SkinThickness e Insulin no subconjunto de dados fornecido têm uma mediana (50%) e quartis (25% e 75%) idênticos (por exemplo, SkinThickness tem Q1=Q2=Q3=29.0 e Insulin tem Q1=Q2=Q3=125.0). Isso é incomum e sugere que uma grande parte dos dados foi imputada com um valor constante (como ZEROs substituídos pela mediana ou média) ou que este subconjunto é altamente concentrado.

Exibir estatísticas descritivas

```
print(df[coluna].describe())
```

Os resultados impressos que demonstram os desbalanceamentos verificados:

Coluna: Pregnancies
count 375.000000
mean 4.333333
std 3.331639
min 0.000000
25% 2.000000
50% 4.000000
75% 7.000000
max 13.000000

Coluna: Glucose
count 375.000000
mean 120.125333
std 28.786375
min 44.000000
25% 101.000000
50% 115.000000
75% 137.000000
max 197.000000

Coluna: BloodPressure
count 375.000000
mean 73.066667
std 10.268759
min 44.000000
25% 66.000000
50% 72.000000
75% 80.000000
max 104.000000
Coluna: SkinThickness
count 375.000000
mean 28.976000

std 4.618257
min 15.000000
25% 29.000000
50% 29.000000
75% 29.000000
max 42.000000

Coluna: Insulin
count 375.000000
mean 124.984000
std 2.213174
min 114.000000
25% 125.000000
50% 125.000000
75% 125.000000
max 135.000000

Coluna: BMI
count 375.00000
mean 31.68240
std 6.22331
min 18.20000
25% 27.30000
50% 31.60000
75% 35.35000
max 49.60000

Coluna: DiabetesPedigreeFunction
count 375.000000
mean 0.387872
std 0.240621
min 0.078000
25% 0.208000
50% 0.300000
75% 0.530500
max 1.191000

Coluna: Age
count 375.000000
mean 34.530667
std 11.622332
min 21.000000
25% 25.000000
50% 31.000000
75% 42.000000
max 66.000000

4. Estratégias de Pré-Processamento e Tratamento de Dados

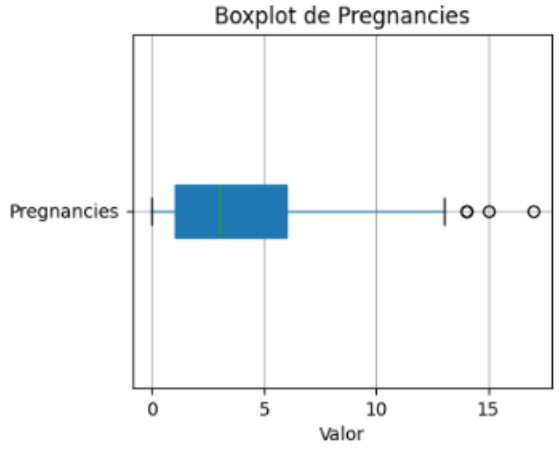
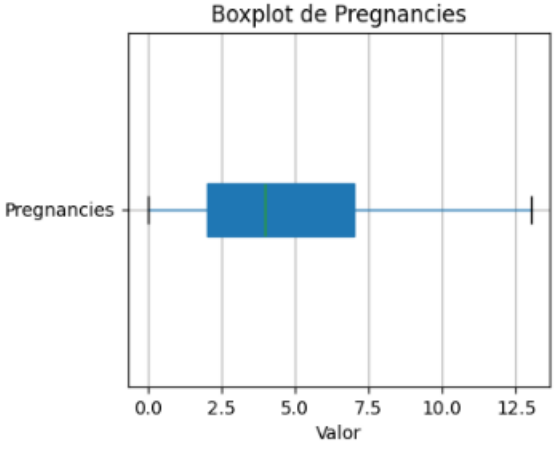
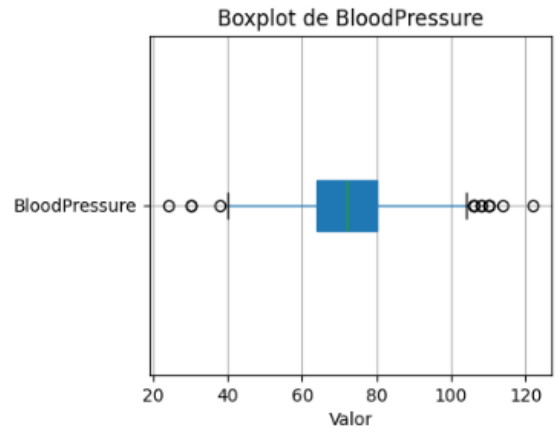
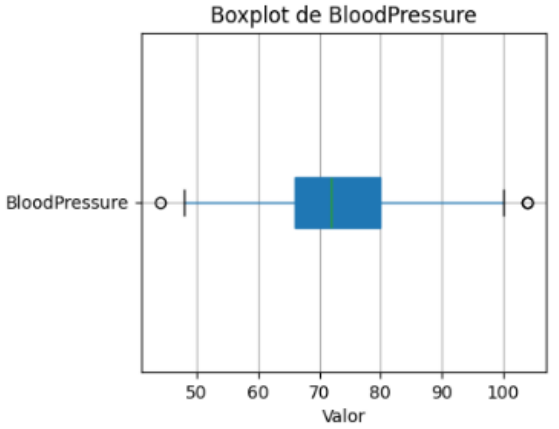
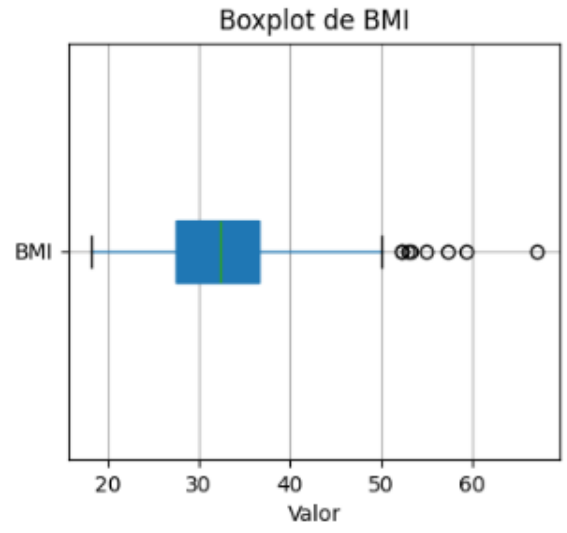
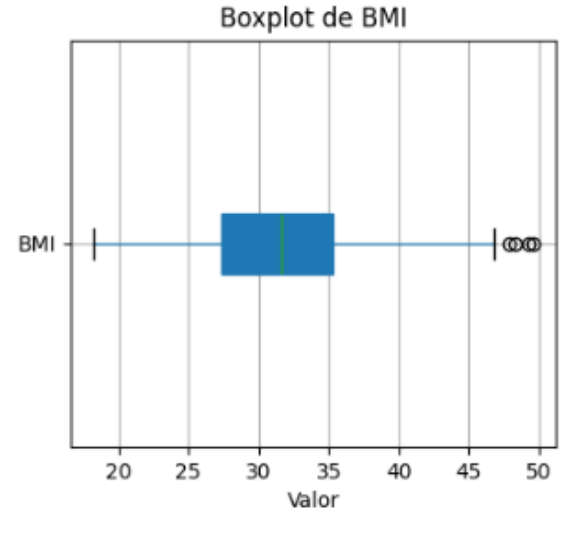
Com base na EDA, as seguintes etapas de tratamento foram realizadas para preparar os dados para a modelagem:

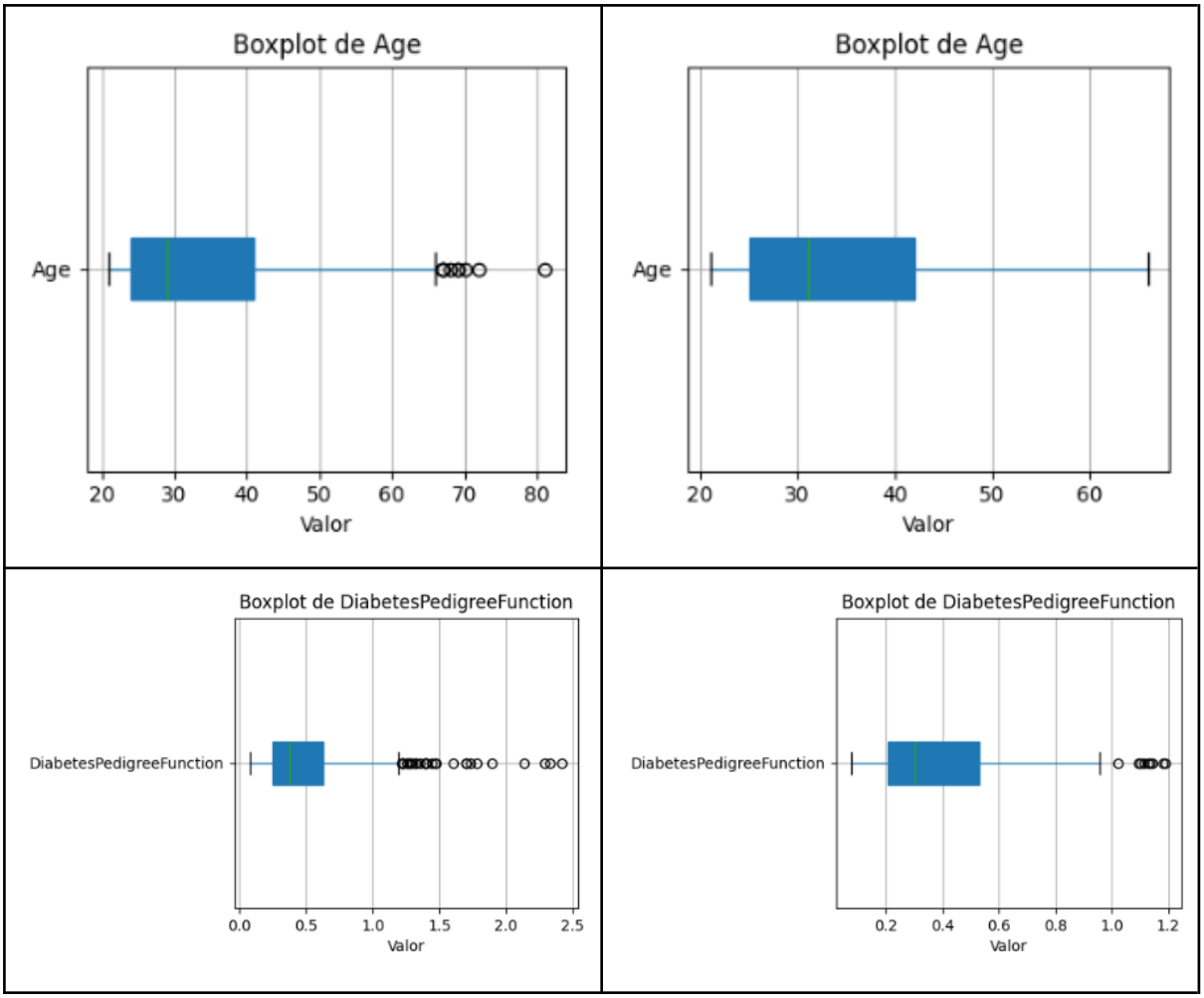
- **Problema 1 - Tratamento de Valores Ausentes:**
  - Identificação de Zeros Improváveis: Valores de 0 são suspeitos e serão substituídos por NaN para que possamos tratá-los corretamente.
  - Os valores zero (0) implícitos nas colunas de Glicose, Pressão Sanguínea, IMC, etc., foram substituídos utilizando **Imputação de Mediana** (ou Média/Moda) para minimizar a distorção introduzida por *outliers*.
  - Contagem de nulos depois do tratamento:
  - ```
    Contagem de valores nulos (NaN) após imputação:
    Pregnancies                 0
    Glucose                     0
    BloodPressure               0
    SkinThickness               0
    Insulin                     0
    BMI                         0
    DiabetesPedigreeFunction    0
    Age                         0
    Outcome                     0
    dtype: int64
    ```
- **Problema 2 - Tratamento dos outliers:**
  - Usando IQR foram detectados outliers nas colunas colunas : ['Pregnancies', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age']
  - Basicamente as colunas que possuem outliers serão tratadas e os dados outliers eliminados do dataset utilizando a técnica **Intervalo Interquartil (IQR)**.
  - O **Intervalo Interquartil (IQR)** é uma técnica estatística robusta e amplamente utilizada para identificar e lidar com *outliers* (valores atípicos) em um conjunto de dados. Ele se baseia na dispersão da metade central dos dados, o que o torna menos suscetível à influência de valores extremos, diferentemente da média e do desvio-padrão.
    ```

Após a o tratamento de outliers, a mudança na contagem de *outliers* nestas colunas é um reflexo direto de como a mediana do conjunto de dados alterou a distribuição e, consequentemente, os limites de 1.5 x IQR

Coluna	Situação ANTES da Alteração (com Zeros)	Situação DEPOIS da Alteração (Zeros → Mediana)
BloodPressure	45 outliers. Muitos zeros (35 casos) estavam presentes e contavam como outliers inferiores (pressão arterial muito baixa, perto de 0).	14 outliers. A substituição dos zeros pela mediana realocou esses 35 pontos para dentro dos limites do IQR, reduzindo drasticamente a contagem de outliers.
SkinThickness	1 outlier. Grande quantidade de zeros (226 casos) tornava a dispersão da coluna enorme, expandindo o IQR e, portanto, o intervalo considerado "normal" ([Q1 - 1.5 x IQR, Q3 + 1.5 x IQR]).	87 outliers. A imputação de 226 zeros pela mediana (29.00) estreitou drasticamente a caixa e as hastes do boxplot (IQR reduzido), tornando o conjunto mais sensível a valores altos e baixos, classificando muitos dados como outliers.
Insulin	34 outliers. Assim como em SkinThickness, o grande volume de zeros (374 casos) aumentou a dispersão e o IQR, camuflando muitos outliers.	346 outliers. A imputação de 374 zeros pela mediana (125.0) concentrou a massa de dados, resultando em um IQR extremamente estreito (5.75). Este IQR minúsculo faz com que quase todo o desvio padrão da coluna seja classificado como outlier, conforme observado na análise detalhada anterior. Isso é um artefato da imputação massiva e não um reflexo da distribuição natural.
BMI	19 outliers. Zeros (11 casos) eram outliers inferiores.	8 outliers. A substituição dos 11 zeros pela mediana (32.0) eliminou os outliers inferiores implausíveis, resultando em uma redução líquida de 11 outliers na cauda inferior. Apenas os outliers superiores (IMC's muito altos) permanecem.

Evidências gráficas da redução ou eliminação de outliers após o tratamento do dado.

Antes	Depois
	
	
	


Boxplot de Age
Age
20
30
40
50
60
70
80
Valor
Boxplot de Age
Age
20
30
40
50
60
Valor
Boxplot de DiabetesPedigreeFunction
DiabetesPedigreeFunction
0.0
0.5
1.0
1.5
2.0
2.5
Valor
Boxplot de DiabetesPedigreeFunction
DiabetesPedigreeFunction
0.2
0.4
0.6
0.8
1.0
1.2
Valor

- **Problema 3 - Normalização/Padronização:**
    - Quando os dados contêm *outliers*, o método de **Padronização (Z-Score Scaling)** é geralmente mais adequado do que a Normalização (Min-Max Scaling).
    - **Padronização (Z-Score):** Transforma os dados para que tenham média = 0 e desvio padrão = 1. É menos sensível a *outliers* extremos, pois os *outliers* simplesmente se tornam valores Z maiores (ou menores), mas não distorcem drasticamente a escala do restante dos dados, como faria o Min-Max.
    - Isso vai garantir que as características sejam igualmente ponderadas sem que os *outliers* exerçam uma influência indevida.
    - O resultado esperado após a padronização é que a média de cada coluna seja muito próxima de 0 e o desvio padrão seja muito próximo de 1. A modificação insere uma etapa que converte o array padronizado de volta para um DataFrame e exibe suas estatísticas descritivas (média, desvio padrão, mínimo e máximo) para confirmação.

5. Modelagem e Motivação do Algoritmo

- **Algoritmo Selecionado:** Foi empregado o modelo **Regressão Logística**.
- **Motivação:** A importância dessa escolha se baseia em dois pontos cruciais: a **natureza do problema** e as **vantagens inerentes** do algoritmo, especialmente a **interpretabilidade**, que é fundamental em contextos de saúde.
- O fator mais importante é que a predição de diabetes, neste dataset, é um problema de Classificação Binária.
- Variável Alvo (Target): A coluna Outcome no seu dataset é a variável que está sendo prevista.
- Resultados Possíveis: Essa coluna possui apenas dois valores:
    - 1: Indica que o paciente tem diabetes (o evento ocorreu).
    - 0: Indica que o paciente não tem diabetes (o evento não ocorreu).
- A Regressão Logística é o modelo de escolha (baseline) para problemas de classificação binária. Ela modela a probabilidade de um paciente pertencer à classe "1" (ter diabetes), ajustando a relação entre as variáveis preditoras (Glicose, IMC, Idade, etc.) e o logaritmo das chances (log-odds) do resultado.

>>>> continuar aqui

6. Treinamento do Modelo

O conjunto de dados foi dividido em subconjuntos de treinamento e teste na proporção de **[Ex: 80% para Treinamento e 20% para Teste]**.

- **Validação:** Foi utilizada a técnica de **Validação Cruzada (Cross-Validation)** com k dobras para garantir que o modelo não estivesse sobreajustado aos dados de treinamento.
- **Otimização (Opcional):** Se realizado, mencionar o ajuste de hiperparâmetros (ex: GridSearchCV, RandomizedSearchCV) para encontrar a melhor configuração do modelo.

7. Análise de Resultados e Desempenho

O modelo treinado foi avaliado utilizando métricas-chave no conjunto de dados de teste, com foco particular na performance da classificação de pacientes diabéticos (classe 1).

Métrica	Valor Obtido (%)	Interpretação
Acurácia	$X.XX\%$	Proporção de predições corretas em geral.
Recall (Sensibilidade)	$Y.YY\%$	Habilidade do modelo em identificar corretamente os casos positivos (evitar Falsos Negativos).
Precisão	$Z.ZZ\%$	Proporção de predições positivas que estavam, de fato, corretas.
F1-Score	$W.WW\%$	Média harmônica entre Precisão e Recall.

A **Matriz de Confusão** demonstrou [Comentar o desempenho do modelo em termos de Falsos Positivos e Falsos Negativos - Ex: "um bom equilíbrio, com um número gerenciável de Falsos Negativos, que é crítico em diagnósticos médicos"].

8. Conclusões e Insights Obtidos

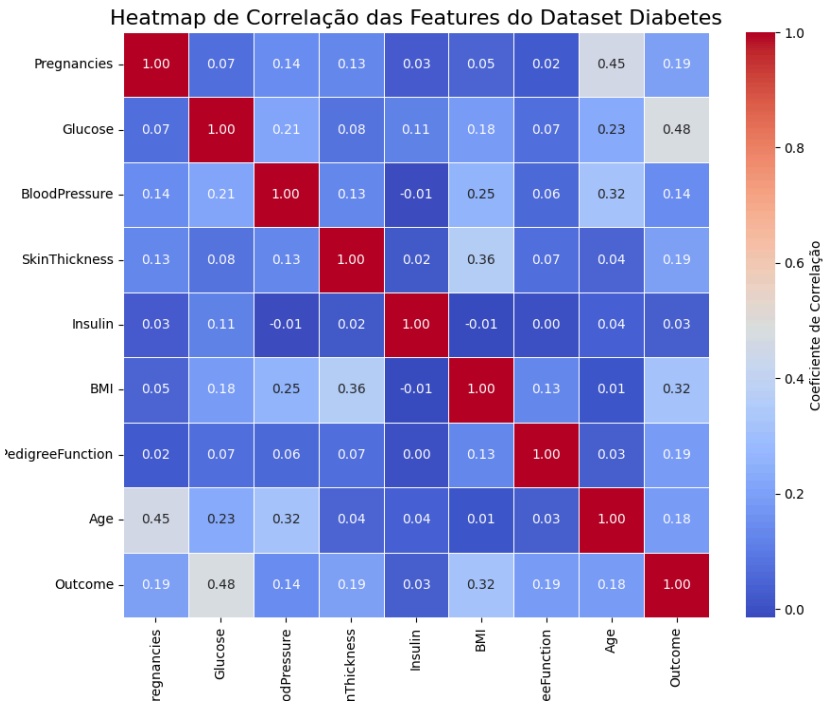
O projeto demonstrou que o modelo **[Nome do Algoritmo]**, após um robusto tratamento de dados, é uma ferramenta promissora para o diagnóstico de diabetes.

- **Insight Principal:** A feature de Glucose foi consistentemente identificada como a mais importante para a predição pelo modelo, reforçando sua relevância clínica.

- **Impacto da Features (Destaque para os 3 maiores):**
    - Os coeficientes mostram a influência de cada *feature* padronizada na probabilidade de ter o **Outcome** 1 (Diabetes):
    - Glucose
        - Coeficiente: 1.128768
    - BMI
        - Coeficiente: 0.588595
    - Pregnancies

        - Coeficiente: 0.389231
    - DiabetesPedigreeFunction
        - Coeficiente: 0.293440
    - SkinThickness
        - Coeficiente: 0.127954
    - Age
        - Coeficiente: 0.081665
    - Insulin
        - Coeficiente: -0.094968
    - BloodPressure
        - Coeficiente: -0.020172

- **Análise de Correção entre as features**
    - Foi aplicada uma análise de correção e foi plotado um gráfico do tipo heatmap



    - A análise de correlação, visualizada através do heatmap, revela insights importantes sobre as relações entre as features e, principalmente, a capacidade preditiva de cada uma em relação ao desfecho (Outcome - Diabetes).
    - Abaixo estão as conclusões resumidas tiradas da análise de correlação:
    - Principais Fatores de Risco (Correlação com o Outcome - diabetes)
    - As features que demonstraram maior correlação com a variável alvo (Outcome) são as mais importantes para a previsão de diabetes:

Feature	Correlação com outcome	Conclusão
Glucose	Forte Correlação Positiva (próximo de 0.50)	É o indicador mais forte de diabetes. Um aumento nos níveis de glicose está altamente associado a um resultado positivo para diabetes.
Pregnancies	Correlação Positiva Moderada (próximo de 0.43)	O número de gestações tem uma influência considerável, indicando que quanto maior o número, maior a probabilidade de um diagnóstico positivo.
Age	Correlação Positiva Moderada (próximo de 0.43)	O número de gestações tem uma influência considerável, indicando que quanto maior o número, maior a probabilidade de um diagnóstico positivo.
BMI (IMC)	Correlação Positiva Moderada (próximo de 0.31)	Um Índice de Massa Corporal mais alto está associado a um maior risco de diabetes.

- Correção entre as próprias features (Risco de Multicolinearidade)
  - Idade e Gravidez (Age e Pregnancies): Existe uma correlação positiva moderada (cerca de 0.51) entre essas duas variáveis, o que é biologicamente esperado (mulheres mais velhas tendem a ter mais gestações). Essa relação não é forte o suficiente para causar problemas graves de multicolinearidade na Regressão Logística, mas é um ponto a ser notado.
  - Outras Correlações Fracas: Após a limpeza e o tratamento de outliers (e imputação da mediana), as correlações entre variáveis que classicamente seriam correlacionadas, como SkinThickness e Insulin, ou BMI e SkinThickness, se tornaram fracas.
- **Conclusão Geral para o Modelo:**
  - A análise de correlação sustenta a ideia de que Glicose, Número de Gestações e Idade são os pilares mais fortes para a construção de um modelo preditivo eficiente para o diagnóstico de diabetes.
  - **Análise de Coeficientes:** O resultado mais valioso no seu código é a **Análise dos Coeficientes** (final do código). O modelo não apenas diz **se** o paciente tem diabetes, mas também **como** cada fator de risco (feature) contribui para essa probabilidade.
  - Os resultados obtidos pelo uso do algorítmo de Regressão Logística mostra que Glucose e BMI (IMC) são as features mais influentes.

Coeficientes positivos altos, como para Glicose, indicam que um aumento nesse valor está fortemente associado a um aumento na probabilidade de ter diabetes.

- **Como pode ser empregado para apoio nos atendimentos:**
    - **Suporte a Decisão Clínica:** Essa interpretabilidade permite que o modelo seja usado como um sistema de **suporte à decisão clínica**. Os médicos podem ver quais fatores estão impulsionando o risco para um paciente individual, não apenas uma caixa preta (como em Redes Neurais mais complexas).