



TechChallenger Fase 1

Trabalho de entrega da Fase 1 da Pós Graduação IA para Devs da FIAP

Integrantes do grupo

- Adalberto Ferreira de Albuquerque Neto (RM368178)
adalbertonet@outlook.com
- Bruno José e Silva (RM367064)
brunojose1977@yahoo.com.br
- Elton de Souza Machado Simão (RM368289)
tonsoumasi@gmail.com
- Lucas Varisco Mendes Bezerra (RM368587)
lucasv.mendes@hotmail.com

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

Sumário

1. Introdução e Objetivos
2. Aquisição e Descrição do Dataset
3. Análise Exploratória de Dados (EDA)
4. Estratégias de Pré-Processamento e Tratamento de Dados
5. Modelagem e Motivação do Algoritmo
6. Treinamento do Modelo e Validação
7. Análise de Resultados e Desempenho
8. Conclusões e Insights Obtidos

1. Introdução, Desafio e Objetivos

Um grande hospital universitário busca implementar um sistema inteligente de suporte ao diagnóstico, capaz de auxiliar 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 um modelo de algoritmo preditivo baseado em Machine Learning para auxiliar no diagnóstico de diabetes.

O objetivo principal consiste no treinamento do algoritmo com dados históricos e prever, com base em medições de diagnóstico, se uma paciente tem diabetes.

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.

Contexto

Este conjunto de dados é de origem do Instituto Nacional de Diabetes e Doenças Digestivas e Renais (National Institute of Diabetes and Digestive and Kidney Diseases).

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/uciml/pima-indians-diabetes-database>

3. Análise Exploratória de Dados (EDA)

5 primeiras 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	

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.
- 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.

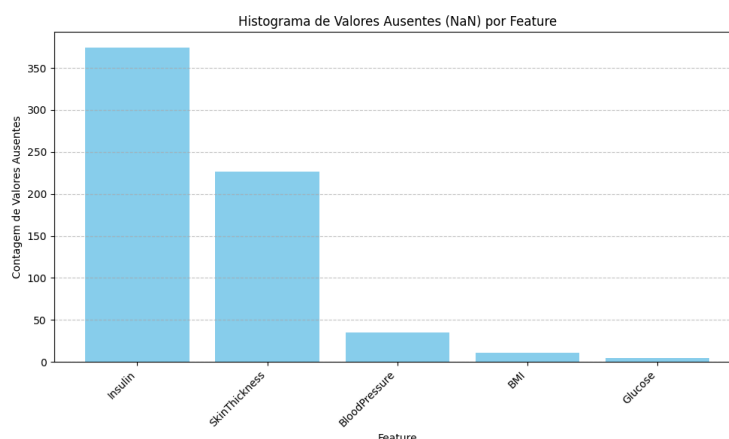
Insight 2

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).

- *Evidências dos zeros anômalos - Histograma*
 - Features com maior quantidade de (NaN);

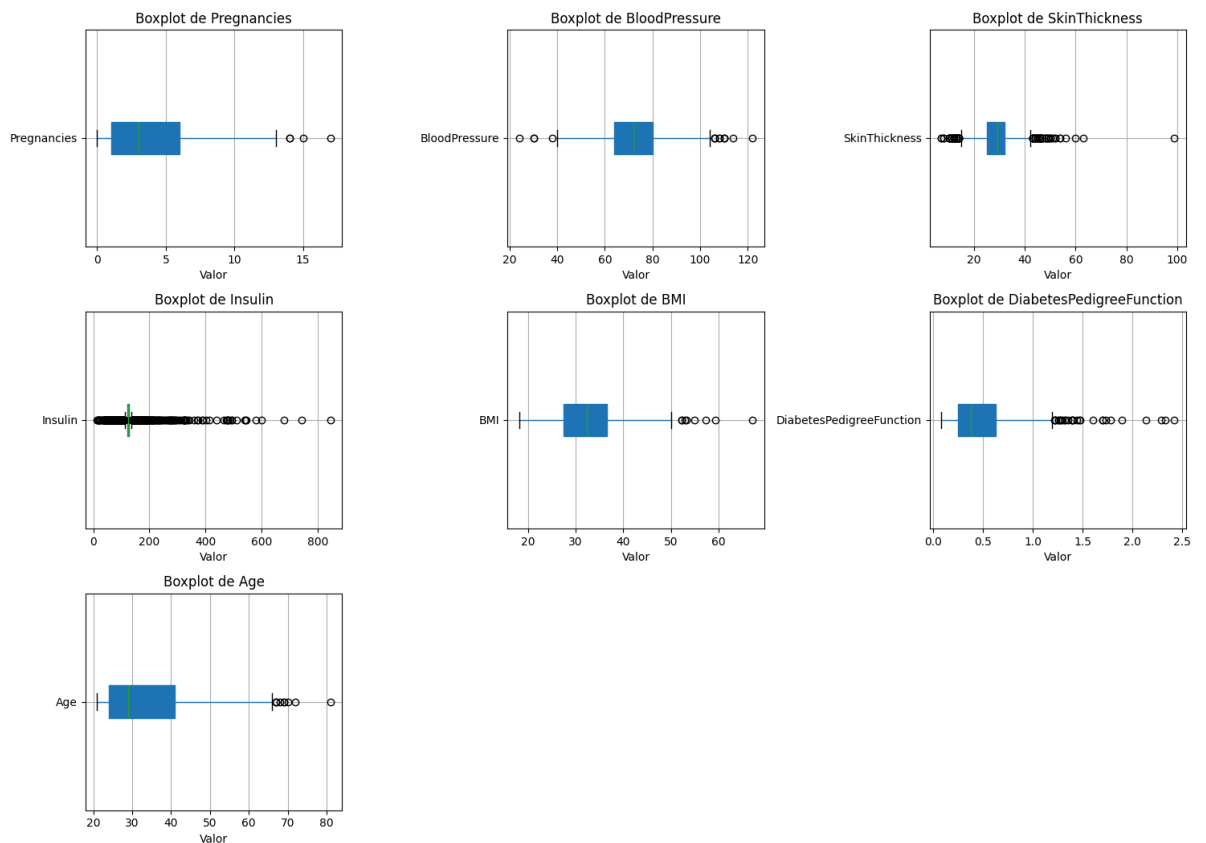


Problema 2 - Detecção de Outliers:

A presença de outliers (valores atípicos) pode prejudicar significativamente o funcionamento de um algoritmo de Regressão Logística, principalmente por afetar a estimativa dos coeficientes do modelo.

Usando IQR foram detectados outliers nas colunas: ['Pregnancies', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age'].

Visualização de Outliers (Método Boxplot) - Dados Após Imputação



Problema 3 - Desbalanceamento:

O desbalanceamento de classes, afeta o funcionamento do algoritmo de Regressão Logística de maneira significativa, gerando um modelo enviesado (tendencioso).

A análise das estatísticas descritivas revela que as diferentes características do dataset estão em escalas variadas e exibem diferentes níveis de dispersão.

- A “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.

Listando as 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 375.000000
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:

Estratégia 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.
- Resultado do tratamento:
 - Os totais verificados NaN por feature

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

- 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*.
- Resultado do tratamento:
 - 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
```

Estratégia Problema 2 - Tratamento dos outliers:

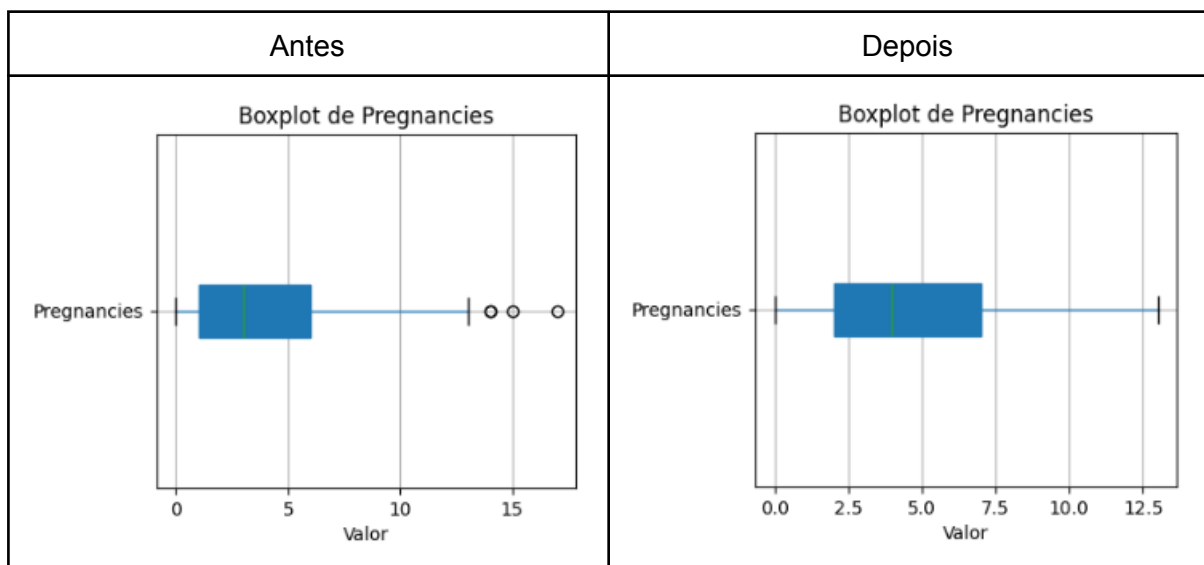
- Usando Intervalo Interquartil (IQR), foram detectados outliers nas colunas : ['Pregnancies', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age'];
- Basicamente as colunas que possuem outliers serão tratadas e os dados outliers eliminados do dataset;
- O Intervalo 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.

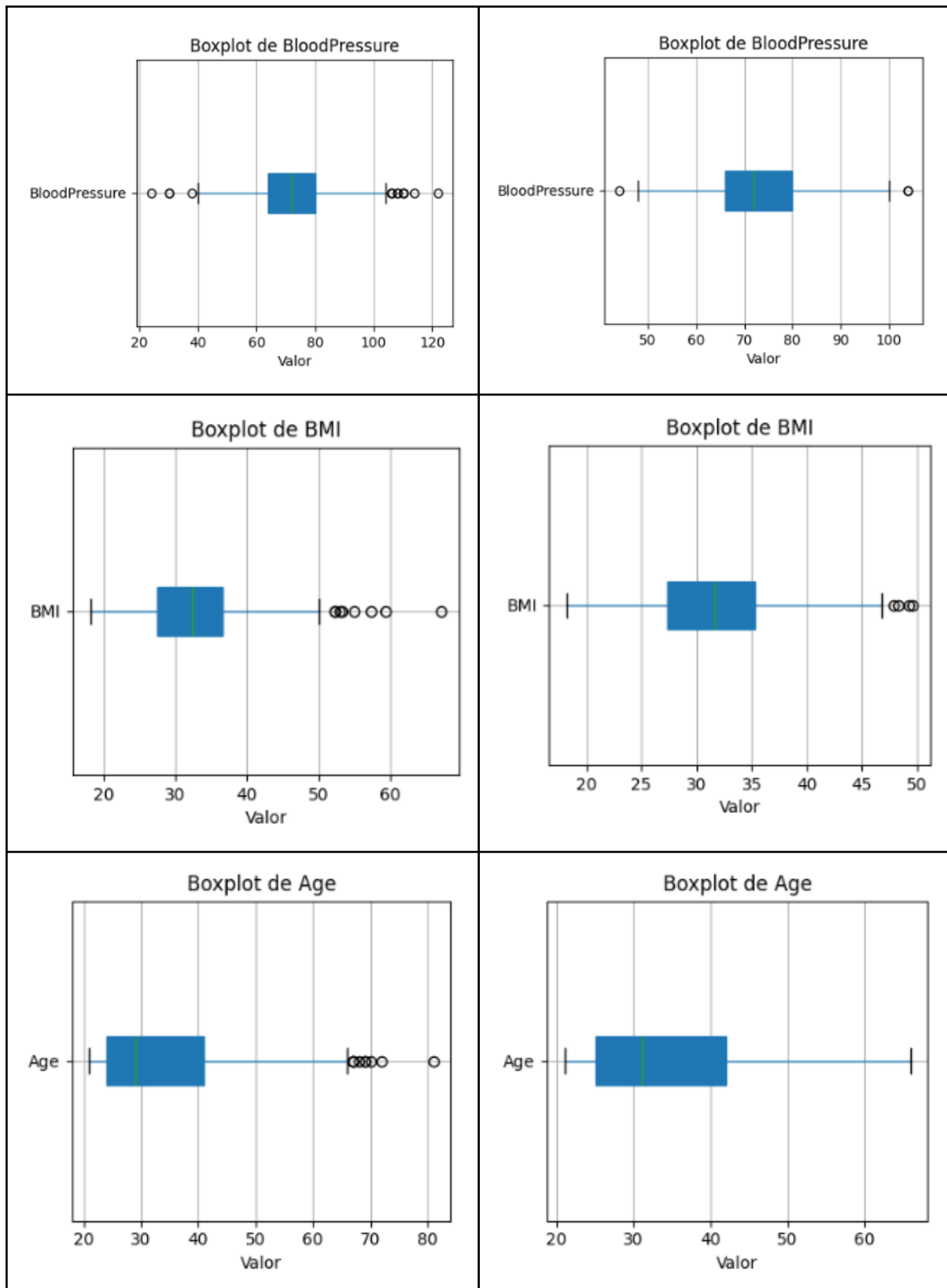
Após a o tratamento de outliers, a mudança na contagem de outliers nestas colunas é um reflexo direto de como a mediana do conjunto de dados alterou a distribuição e, consequentemente, os limites de $1.5 \times \text{IQR}$

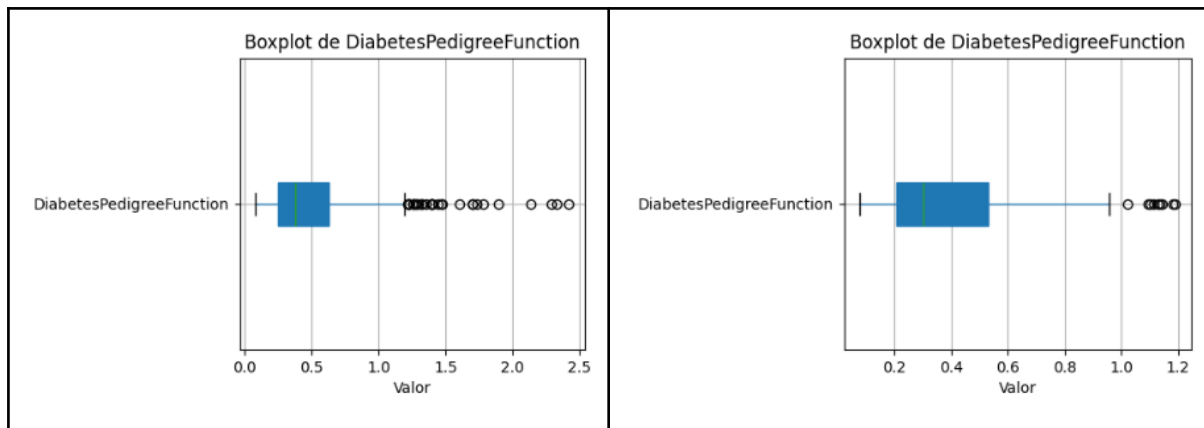
Coluna	Situação ANTES da Alteração (com Zeros)	Situação DEPOIS da Alteração (Zeros → Mediana)
BloodPressure	45 outliers. Muitos zeros (35 casos) estavam presentes e contavam como outliers inferiores (pressão arterial muito baixa, perto de 0).	14 outliers. A substituição dos zeros pela mediana realocou 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" ($[Q1 - 1.5 \times \text{IQR}, Q3 + 1.5 \times \text{IQR}]$).	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 da redução ou eliminação de outliers após o tratamento do dado.







Estratégia 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;

```
model = LogisticRegression(solver='liblinear', random_state=42)
```

- **liblinear** é uma biblioteca otimizada para classificação linear, incluindo Regressão Logística e Máquinas de Vetores de Suporte (SVMs). O solver='liblinear' faz com que o scikit-learn use essa biblioteca específica;

6. Treinamento do Modelo e Validação

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

- Em muitos cenários de saúde, a avaliação de acurácia (`accuracy_score`) de 0.7733\$ (ou 77.33%) pode ser um ponto de partida, mas geralmente não é a métrica mais adequada ou completa para avaliar a *performance* de um modelo;
- **Além disso 77.33% para área de saúde é considerado insuficiente;**

Custos de Erro Diferentes (Falsos Positivos vs. Falsos Negativos)

- Em saúde, os custos dos dois tipos de erro (Falso Positivo e Falso Negativo) são muito diferentes:
- **Falso Negativo (FN):** O modelo diz que o paciente **não tem** a doença, mas ele **tem**.
 - **Custo em Saúde: Alto risco!** Não tratamento, progressão da doença, piora do prognóstico.
 - *Métrica Associada: Recall (Sensibilidade)*
- **Falso Positivo (FP):** O modelo diz que o paciente **tem** a doença, mas ele **não tem**.
 - **Custo em Saúde:** Ansiedade, exames de acompanhamento desnecessários, custos extras.
 - *Métrica Associada: Precision (Precisão)*

7. Análise de Resultados e Desempenho

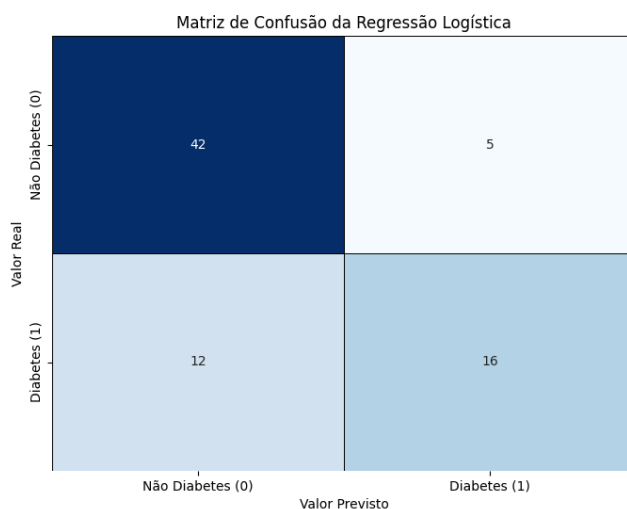
- A métrica que melhor aborda a minimização de **Falsos Negativos (FN)** é o **Recall (Sensibilidade)** para a classe positiva (classe 1 - *Com Diabetes*).
- Com a métrica Recall o resultado foi 0.89 (89%) que geralmente é considerado muito bom e aceitável na maioria dos contextos de Machine Learning.

Comparação de resultados entre diferente métricas

Classe	Precision	Recall	F1-Score	Support
0 (Saudável)	0.78	0.89	0.83	47
1 (Diabétes)	0.76	0.57	0.65	28

A Matriz de Confusão

A Matriz de Confusão foi utilizada para facilitar a análise do desempenho do modelo de classificação dentro do contexto médico.



A matriz de confusão gerada apresentou um excelente resultado fornecendo uma visão clara do desempenho do modelo de Regressão Logística no conjunto de teste.

Valor Real ->	Não Diabético (0)	Diabetes (1)
Não Diabetes (0)	42 (VN) Verdadeiros Negativos	5 (FP) Falsos Positivos
Diabetes (1)	12 (FN) Falsos Negativos	16 (VP) Verdadeiros Positivos

8. Conclusões e Insights Obtidos

O projeto demonstrou que o modelo **Regressão Logística**, 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.

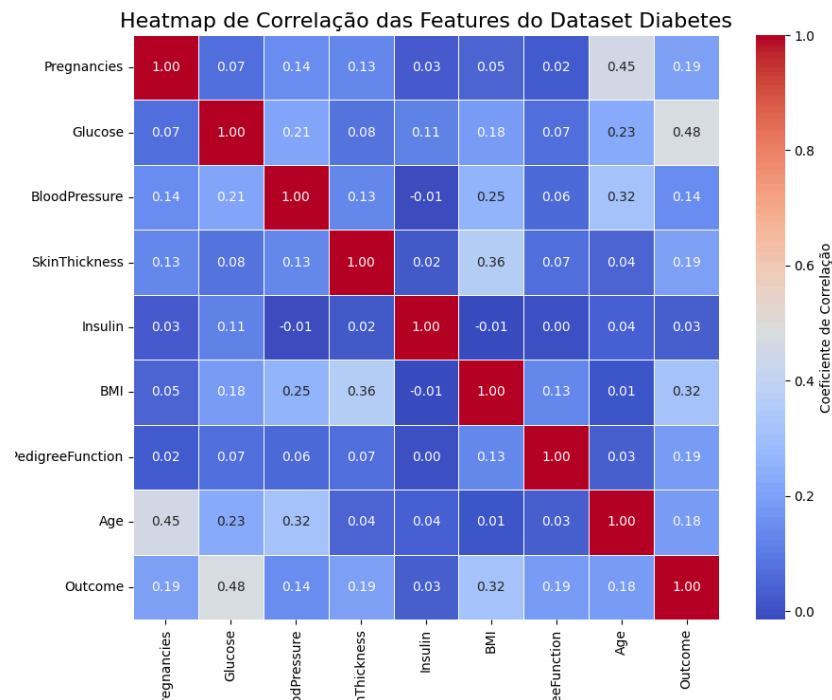
Impacto da Features (Destaque para os 3 maiores):

- Os coeficientes mostram a influência de cada *feature* padronizada na probabilidade de ter o **Outcome 1** (Diabetes):

Feature	Coeficiente:
Glucose	1.128768
BMI	0.588595
Pregnancies	0.389231
DiabetesPedigreeFunction	0.293440
SkinThickness	0.127954
Age	0.081665
Insulin	-0.094968
BloodPressure	-0.020172

Análise de Correção entre as features

- Visualização gráfica da correlação entre as features em um gráfico tipo Mapa de Calor (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).

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)	<p>Direção da Relação (Positiva): O sinal positivo indica que, em geral, existe uma tendência de que à medida que a idade ("Age") aumenta, a probabilidade ou a ocorrência de um resultado positivo para diabetes (outcome = 1) também tende a aumentar.</p> <p>Força da Relação (Moderada): O valor de 0.43 (que está entre 0.3 e 0.7 ou próximo disso) sugere uma relação moderada.</p> <p>Isso significa que a idade é um fator importante, mas não é o único e nem o mais forte preditor do resultado.</p> <p>A correlação não é fraca (abaixo de 0.3), mas também não é forte (acima de 0.7).</p>
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) -

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 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 algoritmo de Regressão Logística demonstram 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 à 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).