📄 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

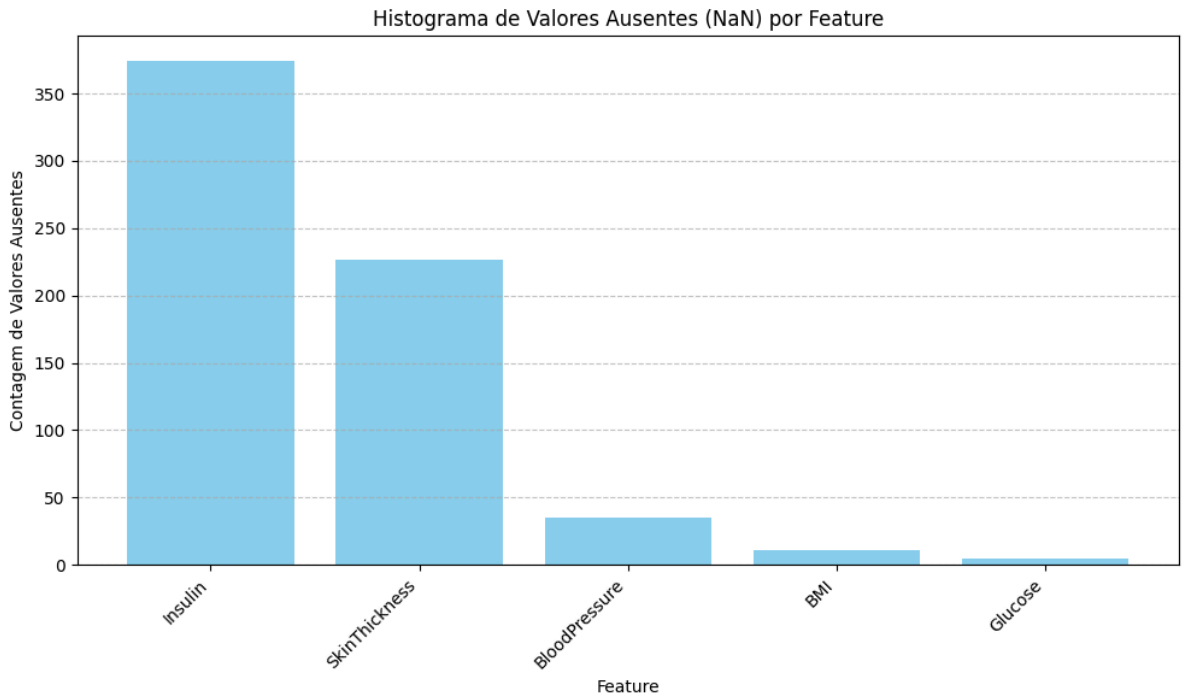
Insulin (insulina)

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.

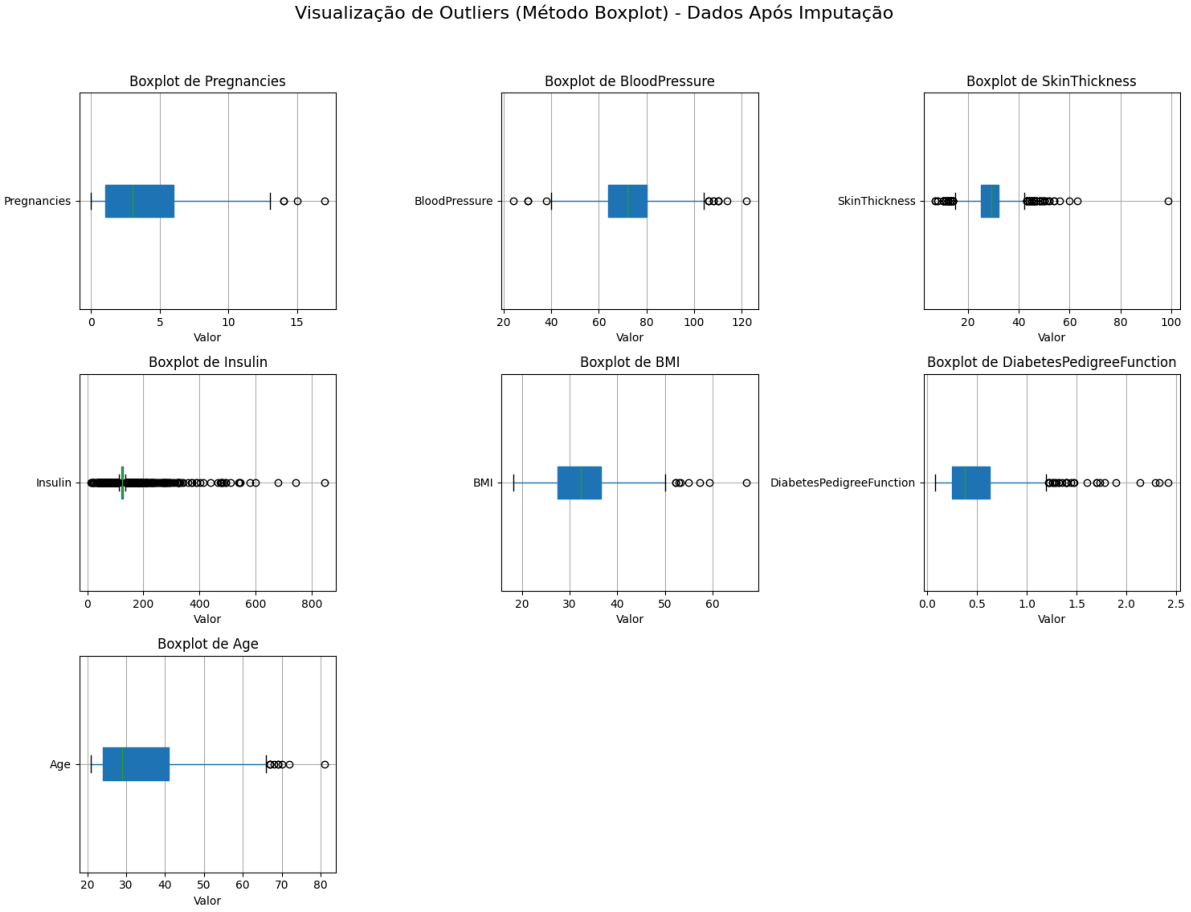
A forma como essa relação se manifesta difere entre os principais tipos de diabetes: Tipo 1 (DM1) e Tipo 2 (DM2)

```
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:** Verificou-se um desbalanceamento na variável alvo, com a classe 'Não-Diabético' sendo significativamente mais representada do que a classe 'Diabético'.
    - *Visualização:* Gráfico de setores ou contagem das classes.
- **Problema 4 - Outliers:** A presença de *outliers* em algumas variáveis (ex: Insulina e Pedigree Function) foi identificada por meio de *boxplots*, o que poderia afetar a robustez do modelo.

4. Estratégias de Pré-Processamento e Tratamento de Dados

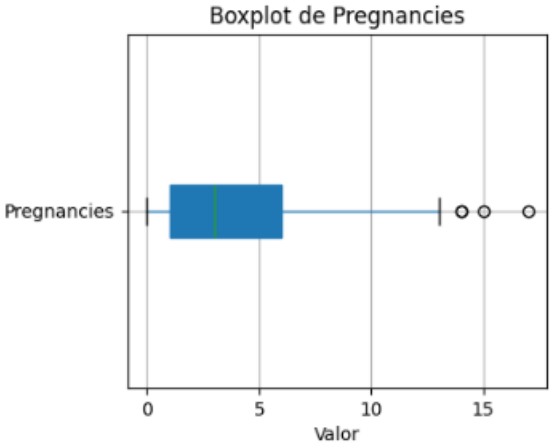
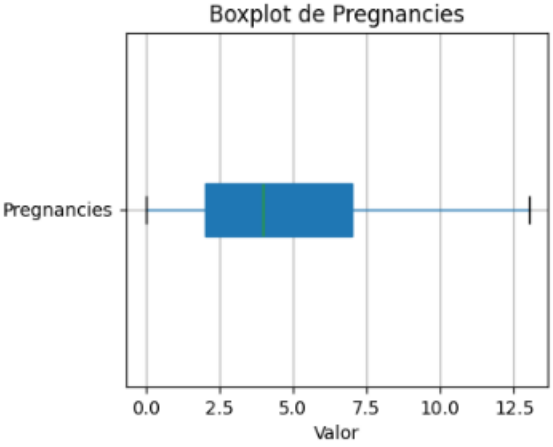
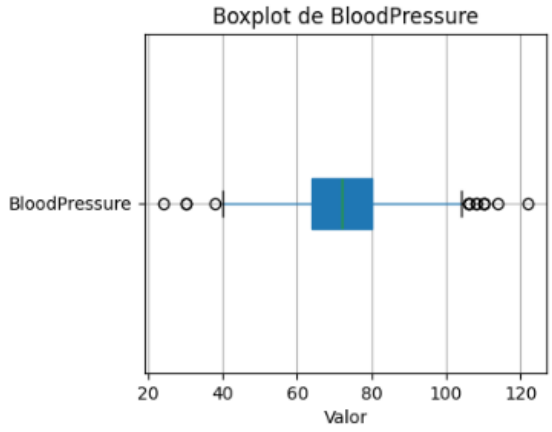
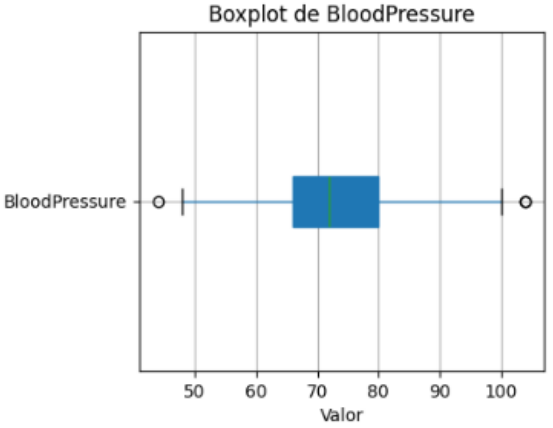
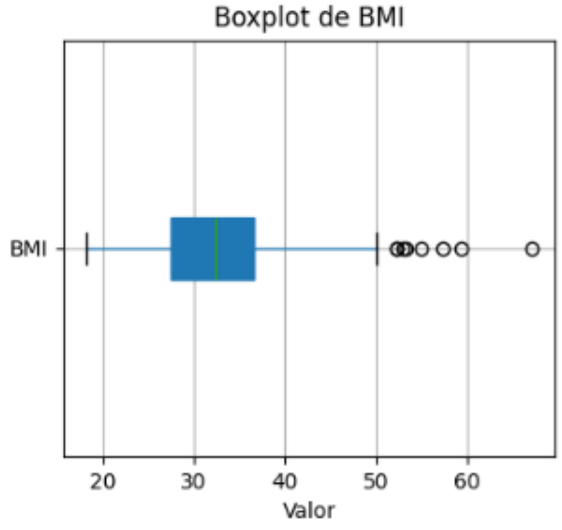
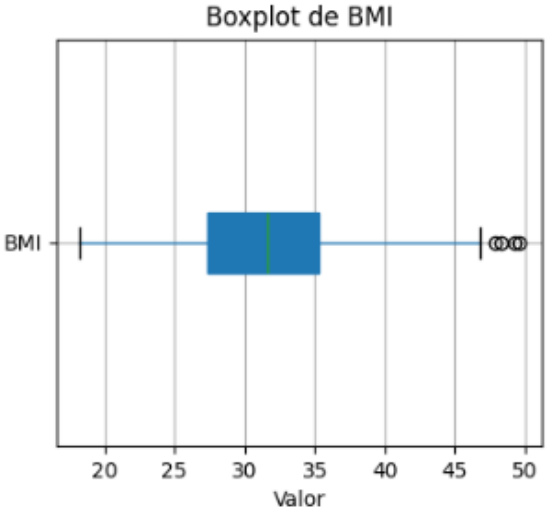
Com base na EDA, as seguintes etapas de tratamento foram realizadas para preparar os dados para a modelagem:

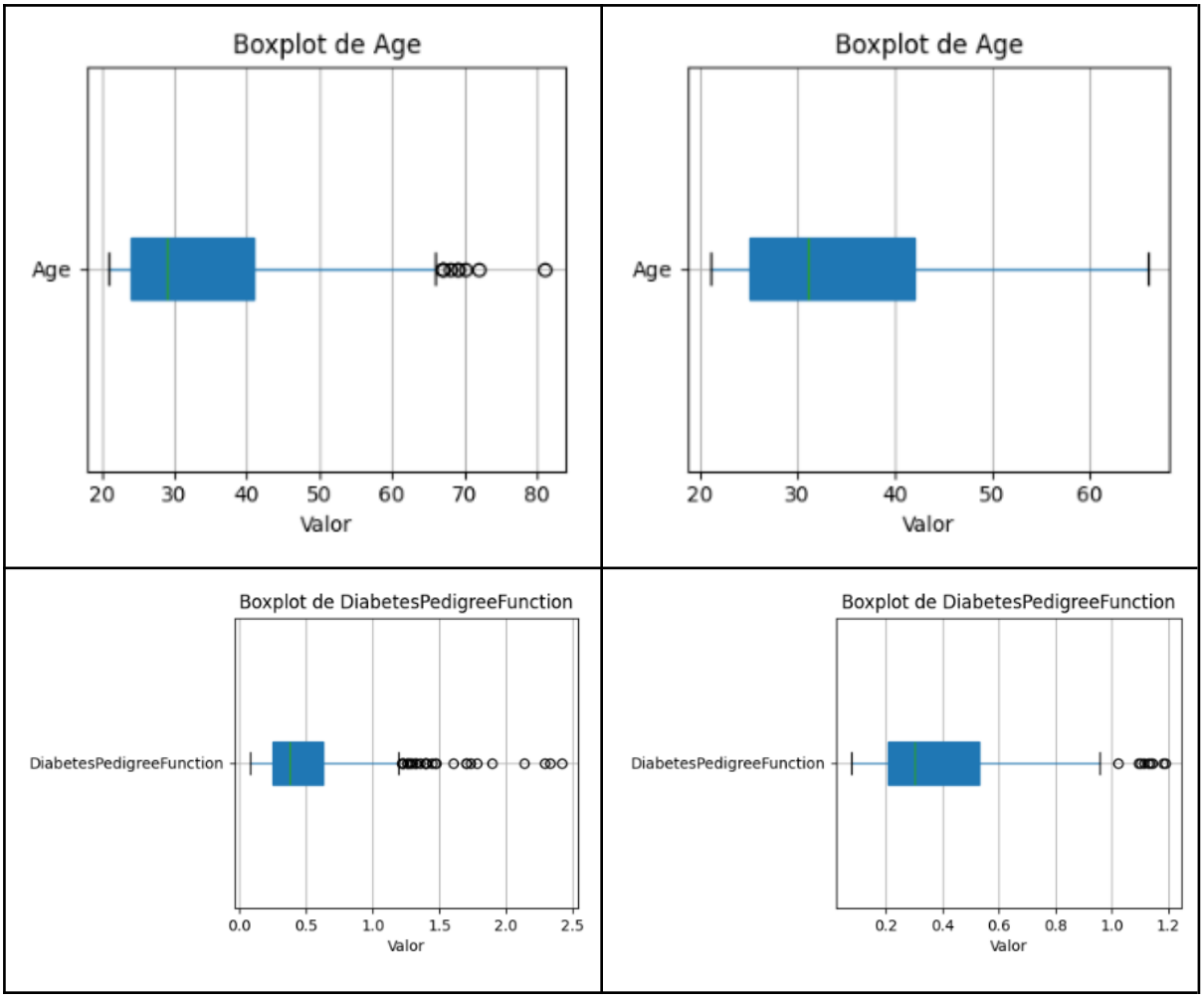
- **Tratamento de Valores Ausentes:**
    - Identificação de Zeros Improváveis: Valores de 0 são suspeitos e serão substituídos por NaN para que possamos tratá-los corretamente.
    - Os valores zero (0) implícitos nas colunas de Glicose, Pressão Sanguínea, IMC, etc., foram substituídos utilizando **Imputação 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
      ```
- **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.

Evidencias 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

>>>> continuar aqui

- **Normalização/Padronização:** As variáveis numéricas foram **[Padronizadas (StandardScaler) ou Normalizadas (MinMaxScaler)]** para garantir que todas as características contribuam igualmente para o treinamento do modelo.
- **Tratamento de Desbalanceamento (Opcional):** Se realizado, mencionar a técnica (ex: SMOTE, *Under-sampling* ou uso de pesos de classe).

5. Modelagem e Motivação do Algoritmo

- **Algoritmo Selecionado:** Foi empregado o modelo **[Nome do Algoritmo - Ex: Random Forest Classifier]**.
- **Motivação:** A escolha deste algoritmo deve-se à sua **[Indicar a Razão - Ex: alta robustez contra *overfitting*, capacidade de lidar com a não-linearidade dos dados e facilidade de interpretar a importância das features]**. Para a tarefa de classificação binária, o modelo oferece um balanço eficaz entre complexidade e desempenho preditivo.

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 **[Nome da Feature - Ex: Concentração de Glicose ou IMC]** foi consistentemente identificada como a mais importante para a predição pelo modelo, reforçando sua relevância clínica.
- **Próximos Passos:** Sugestões para melhorias futuras incluem a exploração de modelos de *Ensemble* mais complexos ou a coleta de dados adicionais para mitigar o desbalanceamento inicial.

Gostaria de ajuda para detalhar o conteúdo técnico de algum desses capítulos (ex: quais códigos mostrar, quais gráficos incluir) para o seu vídeo ou para a documentação?