UMA ABORDAGEM COM MODELOS DE APRENDIZADO DE MÁQUINA

PROF. DR. MURILO COELHO NALDI

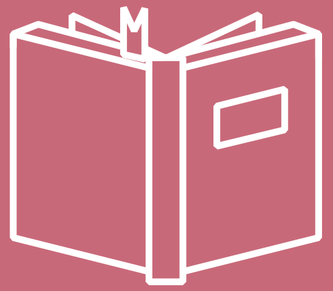
BRUNO LEANDRO PEREIRA
RA 791067

CARLOS EDUARDO FONTANELI
RA 769949

IVAN DUARTE CALVO
RA 790739

Problemática

Escolha e estudo de conjunto de dados, gerar e avaliar um modelo classificador



Conjunto de Dados

O dataset escolhido possui o objetivo de classificar o remédio com a melhor resposta utilizado para o tratamento des pacientes com uma determinada doença.

Atributos

- Idade
- Sexo
- Pressão Sanguínea
- Nível de Colesterol
- Relação Sódio/Potássio
- Rémedio Indicado(Alvo)

Estudo dos Dados

Realização da análise descritiva e exploratória dos dados.

Treino, Teste e Avaliação de Modelos

Implementação dos modelos de DECISION TREE, KNN E GAUSSIAN NAYVE BAYES, com treino, teste e métricas de avaliação.

Objetivos & Metodologia

Objetivos

Gerar modelos preditivos e analisar o desempenho dos mesmo para o conjunto de testes. Ademais, buscou-se inferir as qualidades e defeitos de cada modelo.

Metodologia

Análise exploratória dos dados.

Implementação sistematizada de modelos, com treino e teste sobre o conjunto de dados.

Avaliação dos resultados obtidos.

Ferramentas

Linguagem de programação multi-paradigma orientada a objetos: **Python**.

Jupiter Notebooks para realização do relatório.

Canva para elaboração da apresentação.

Análise Exploratória

Obtenção de informações mais palpáveis e interpretativas

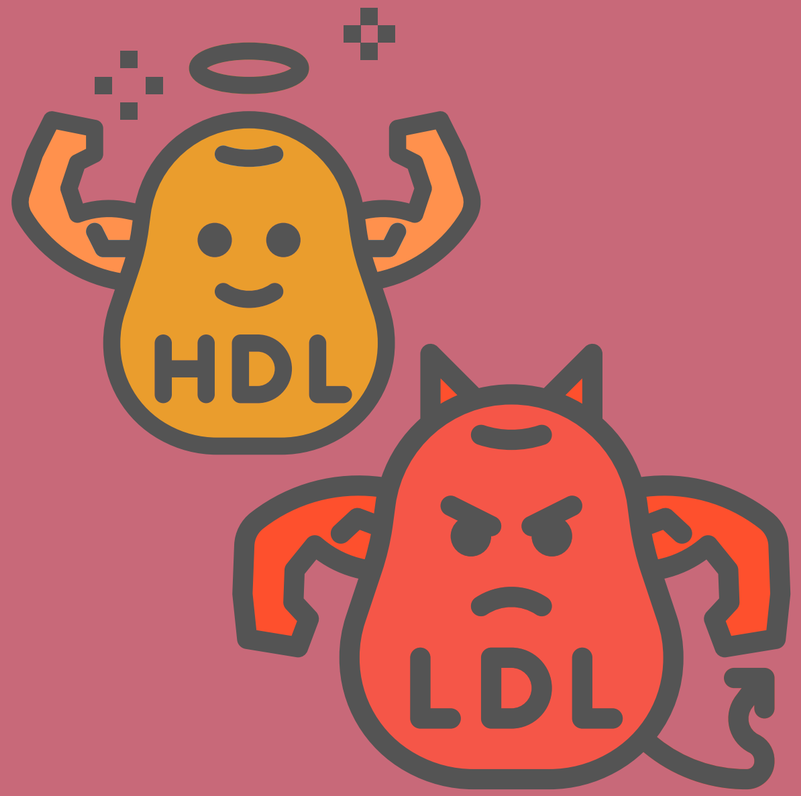
Busca de possíveis **outliers** e/ou desbalanceamentos

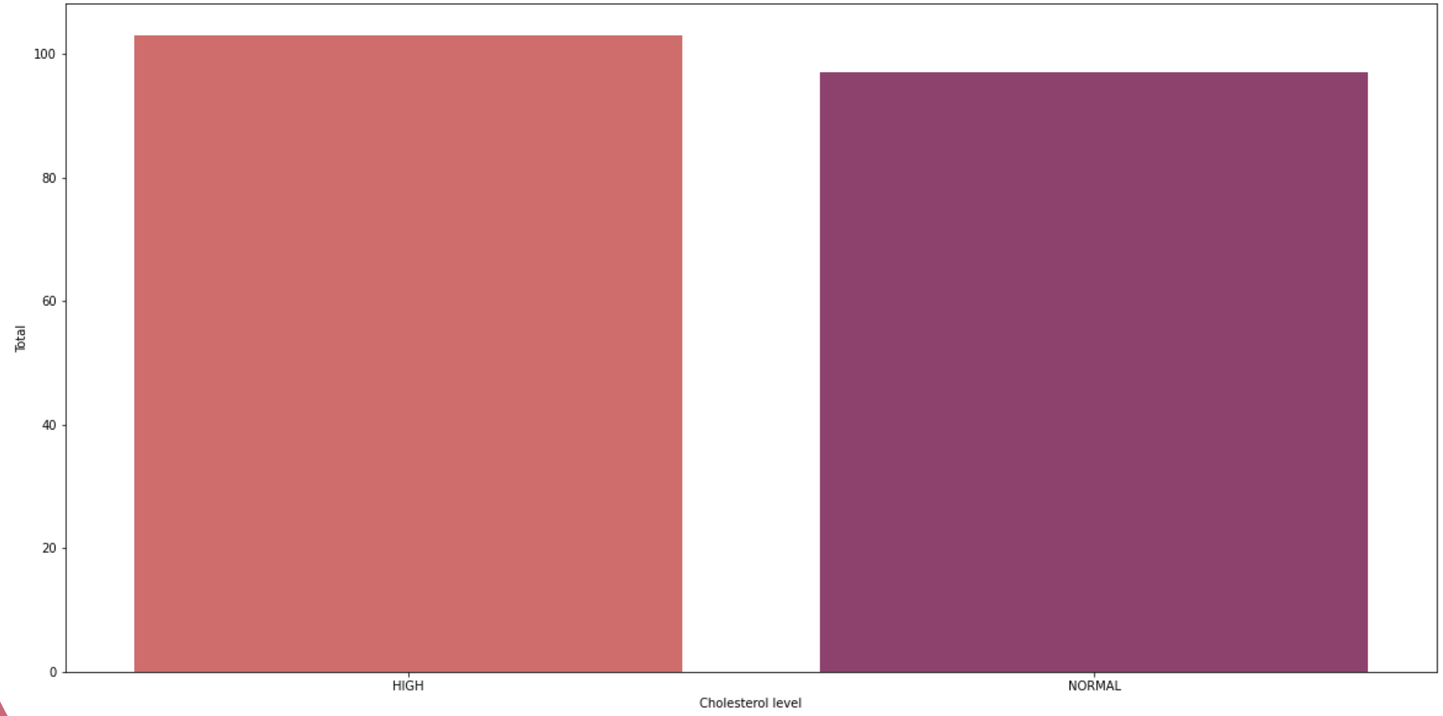
Representação visual dos dados através de gráficos

Boxplot das variáveis para representação de um conjunto de observações de uma variável quantitativa

Colesterol

- Distribuição equilibrada
- Recorrência um pouco maior na classe alto(high)

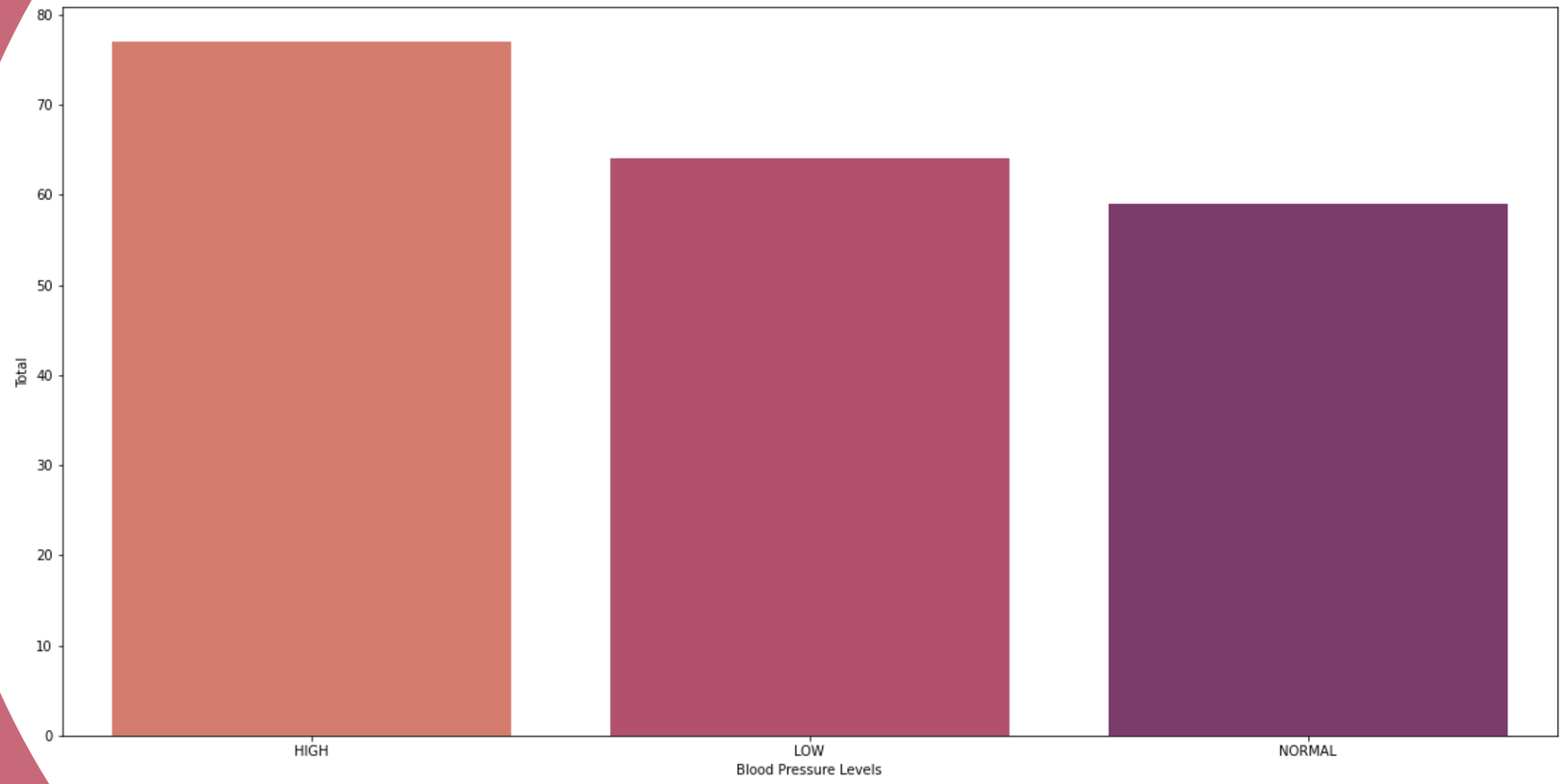




Nível de colesterol (alto ou baixo)

Pressão Sanguínea

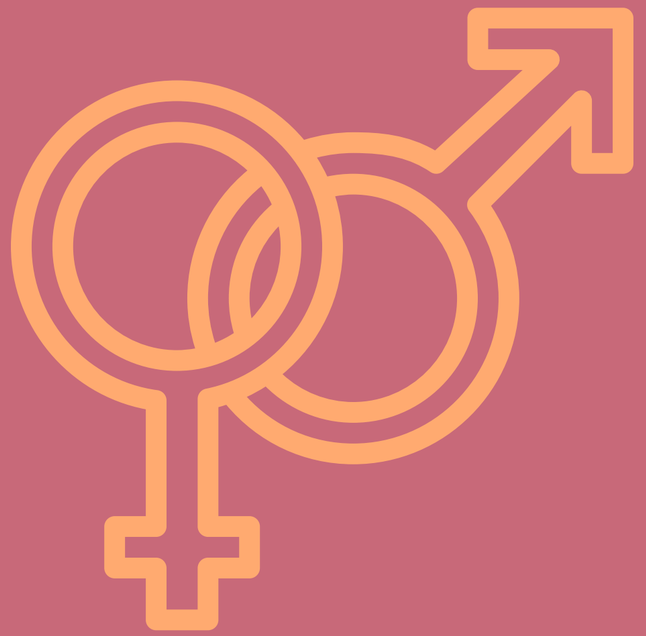
- Distribuição relativamente equilibrada
- Maior recorrência na classe alta(high) seguida da baixa(low)
- Condizente com o contexto de pacientes doentes

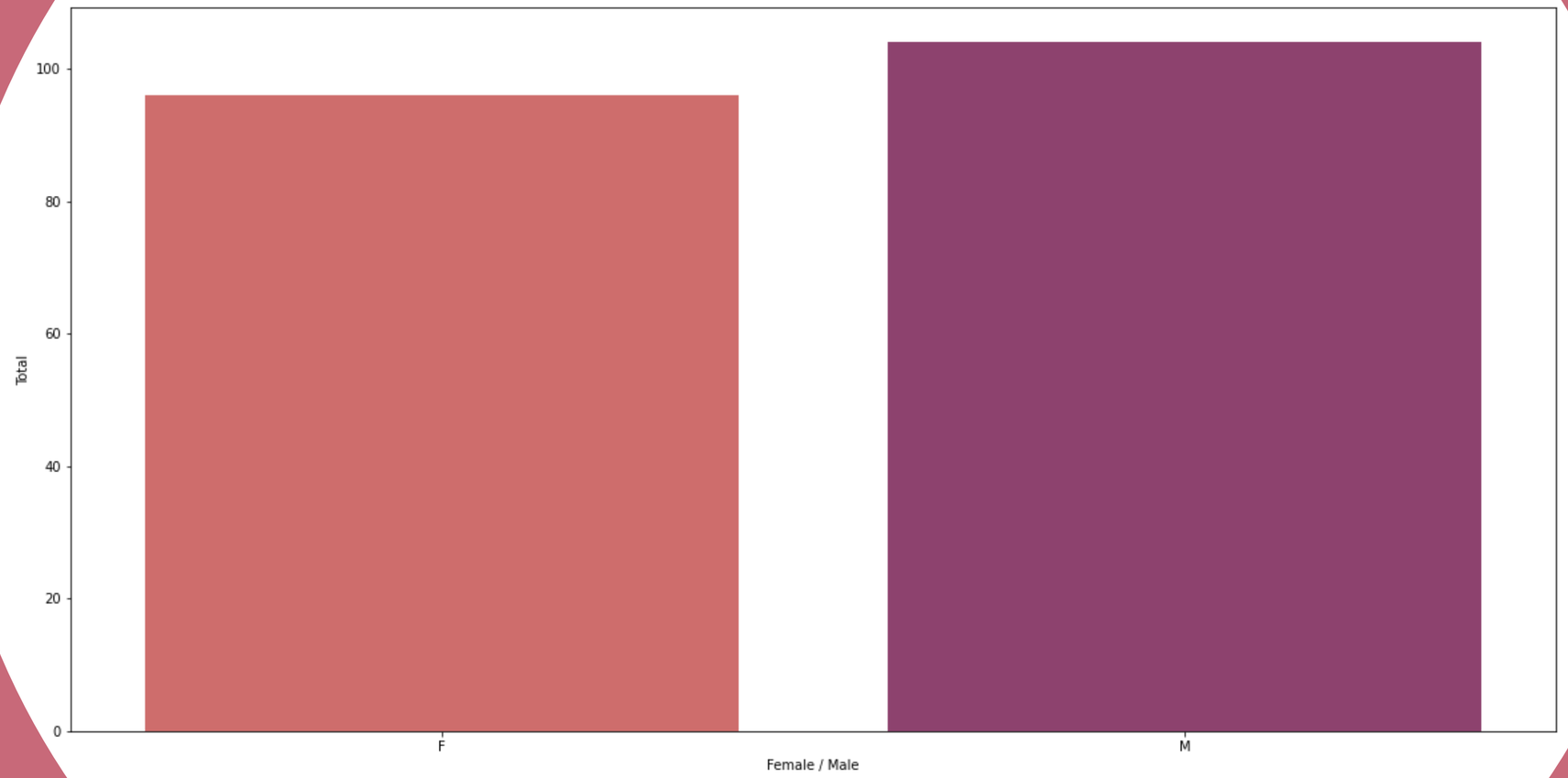


Pressão sanguínea (baixo, normal ou alto)

Sexo

- Distribuição equilibrada
- Recorrência um pouco maior na classe masculino(M)

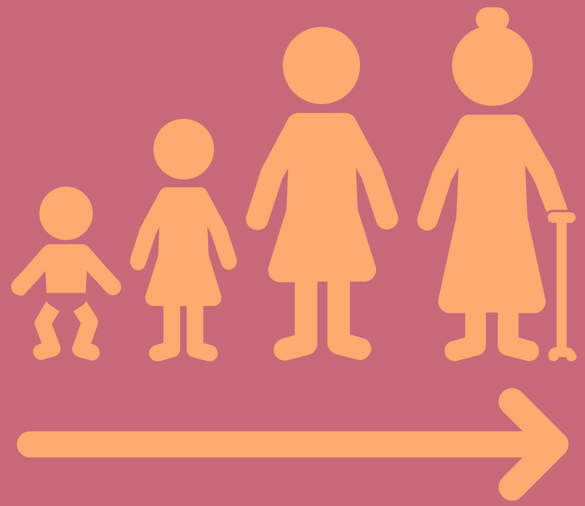


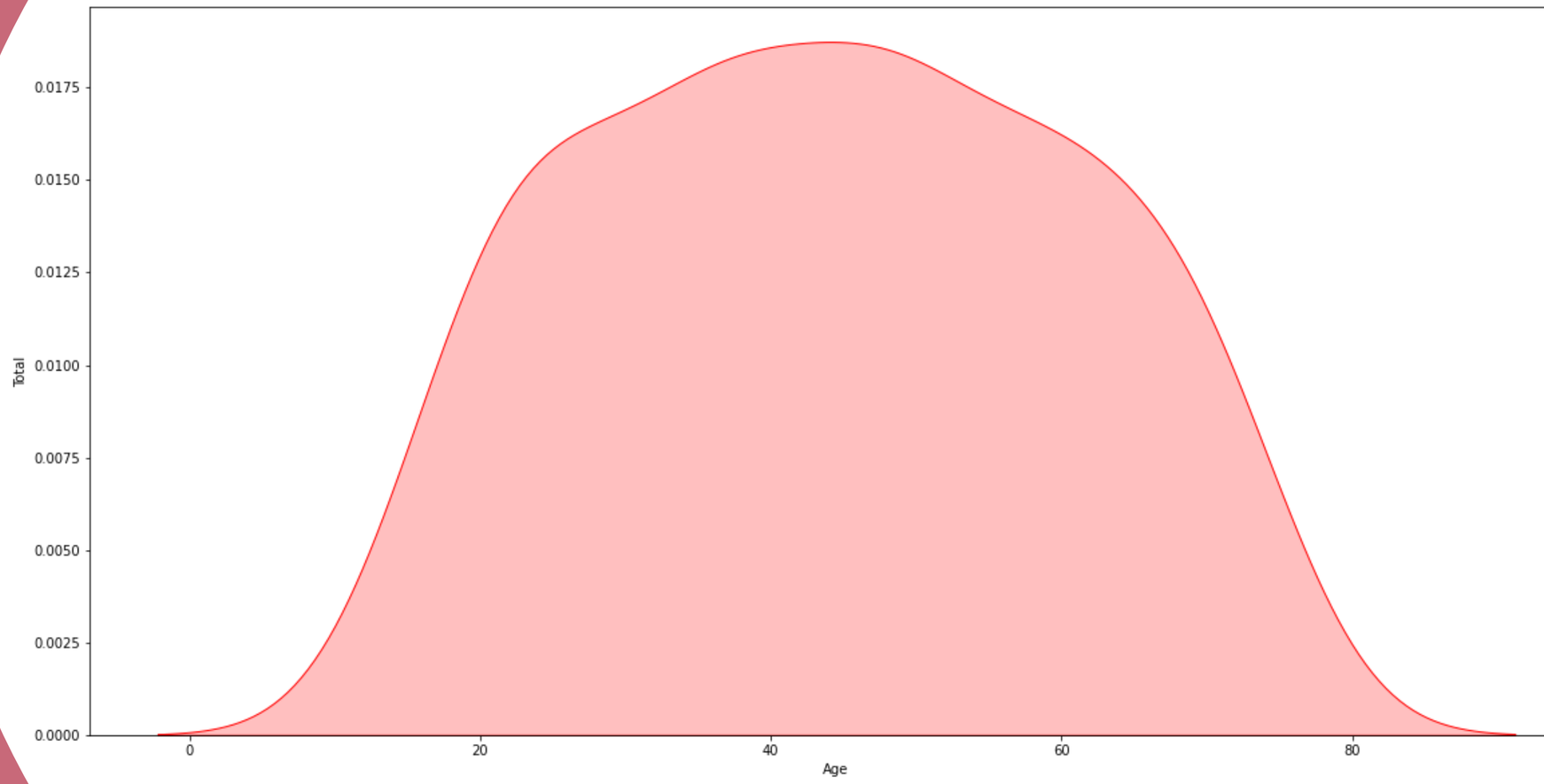


Sexo (feminino ou masculino)

Idade

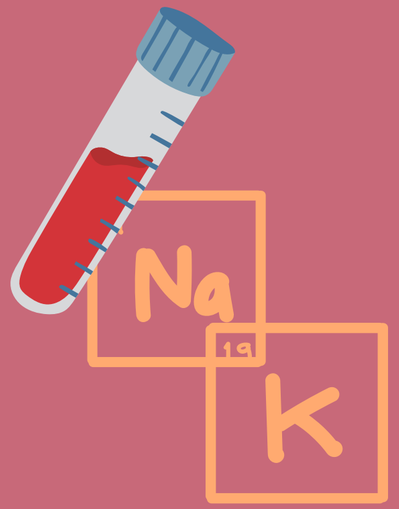
- Maiores recorrências tendem a ser ao redor da média, similar a uma distribuição normal
- Comportamento condizente visto que o atibuto medida um fenômeno natural

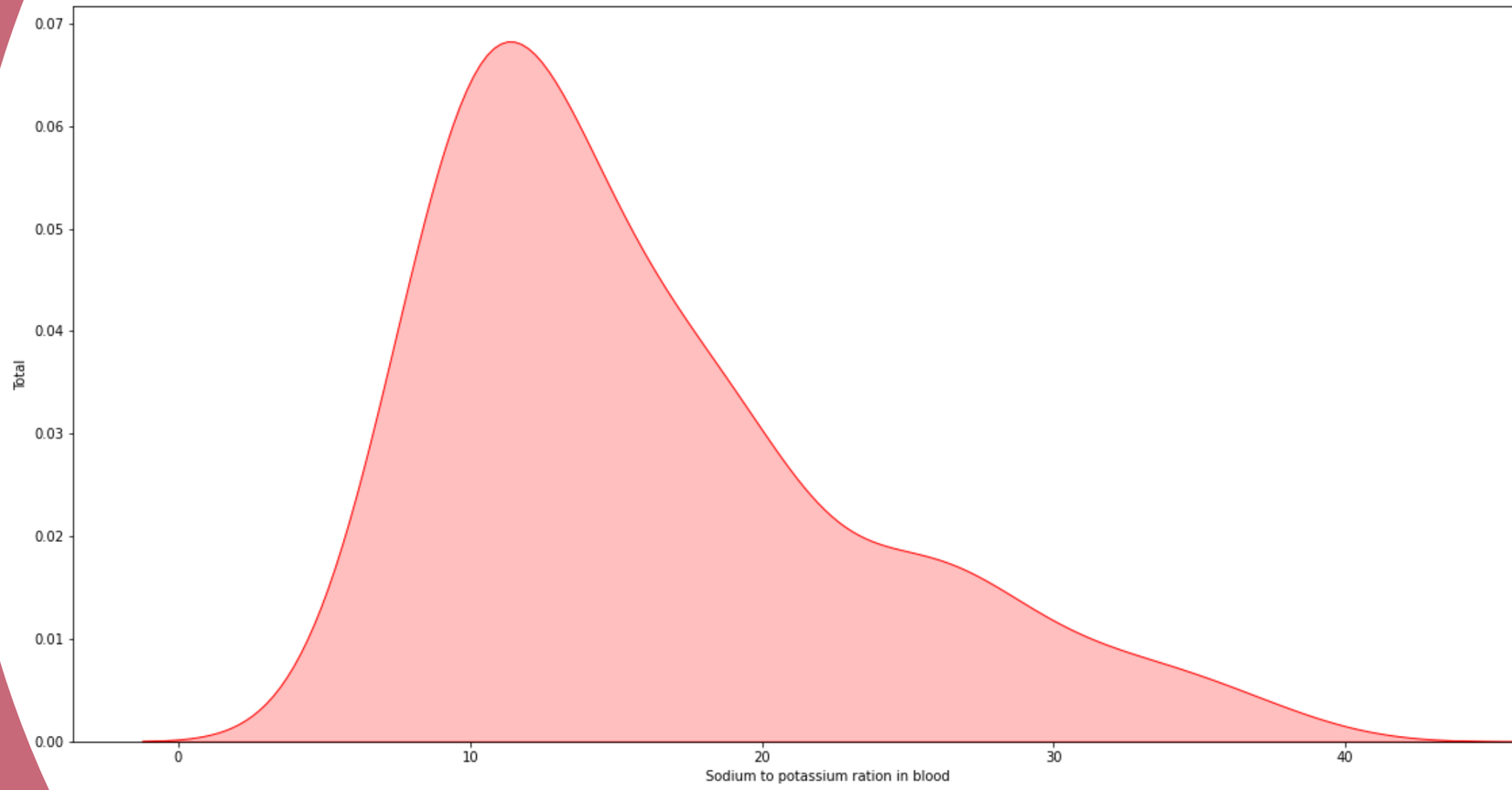




Proporção de Sódio para Potássio no Sangue

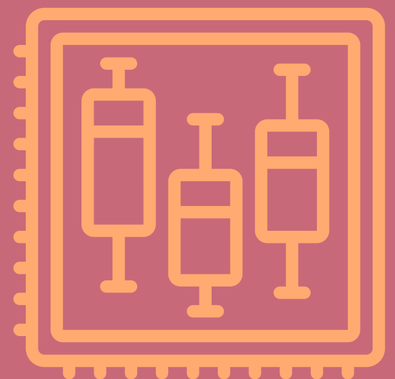
- **Predominância nos níveis próximos de 10, o que pode-se ser um resultado esperado para pessoas com alguma irregularidade dado que níveis recomendados de proporção são de 1:3**



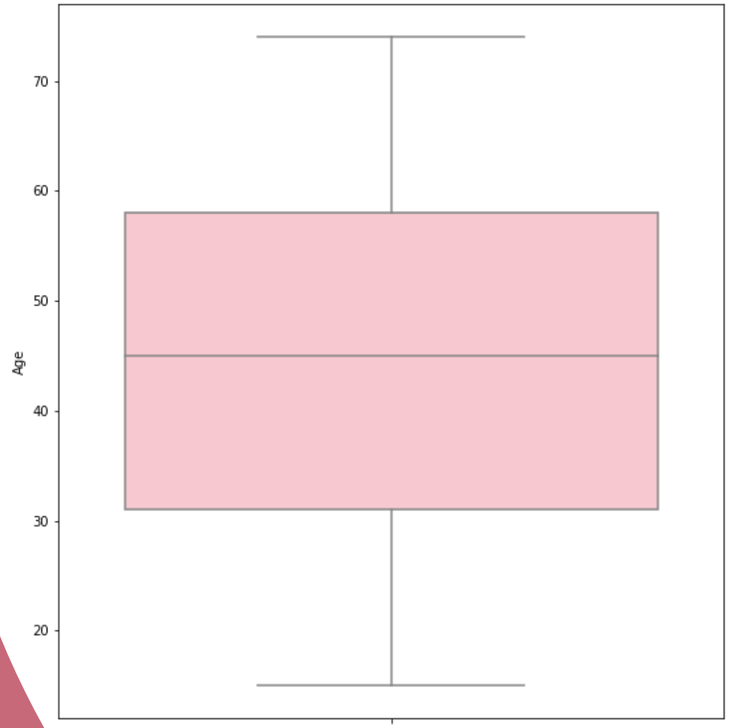


Boxplots das variáveis contínuas e discretas

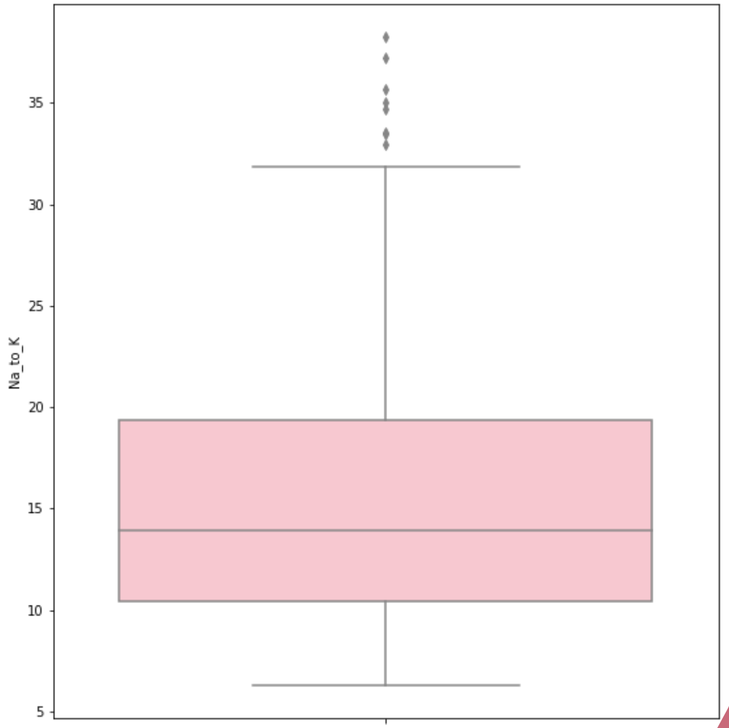
- Sem outliers em Idade, possivelmente relacionado ao provável comportamento de distribuição normal
- Possíveis outliers em Sódio-Potássio
- Alguns valores acima de 32%;



Idade

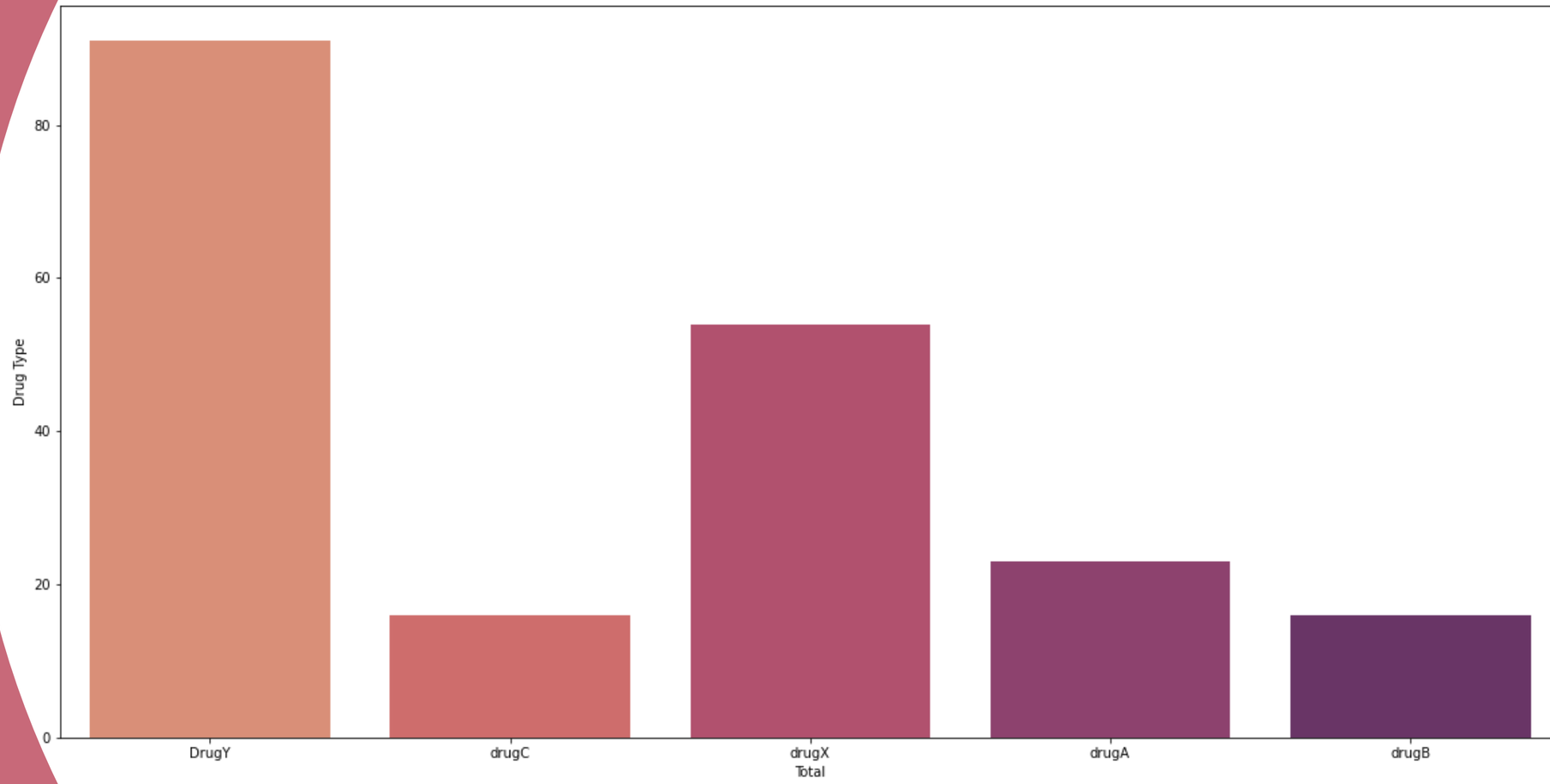


Sódio-Potássio



Remédio (alvo)

- Classes A, B e C com significativamente menos recorrência;
- Classe Y com recorrência 4x maior que outras classes;



Remédio (A, B, C, X ou Y)

Pré-Processamento Base

Separação entre atributos e alvo;

Discretização dos atributos nominais ([0, número_de_classes - 1]);

Divisão entre treino(70%) e teste(30%);

Normalização dos atributos contínuos para KNN(cal. distancia) e GaussNB(distribuição gaussiana);

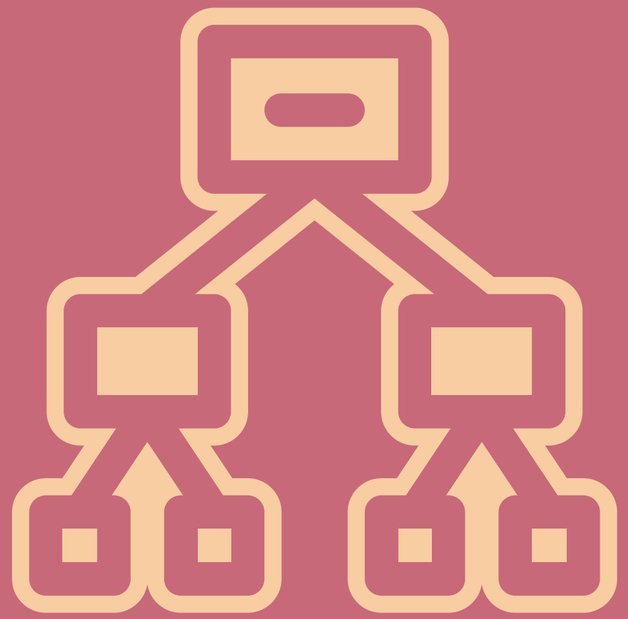
Pré-Processamento Adicional

Processo opcional, usado para testagem se houve ou não ganho/perda de desempenho e se o mesmo vale a pena.

Remoção de outliers, exclusão de tuplas com valore muito discrepantes para o atributo Na_to_K

Duplicação de tuplas com classes alvos menos recorrentes para balancemanto dos dados;

Árvore de Decisão



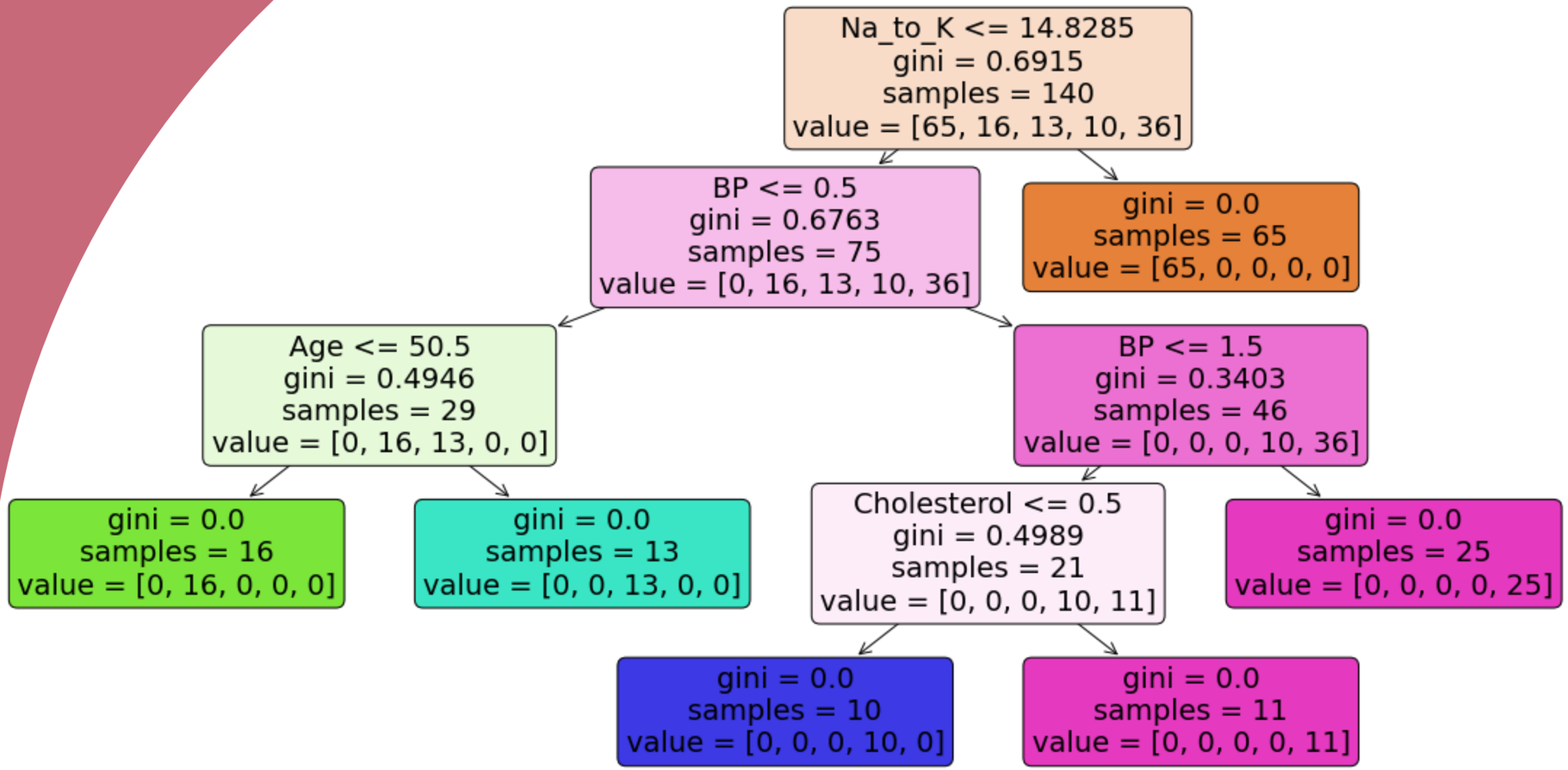
Divisão do conjunto de dados baseados em condições

Treinamento, teste e avaliação

Cálculo de métricas de avaliação

Análise visual dos resultadoas com matriz de confusão

Árvore Gerada



Ex. Gini

- De acordo com o Indice Gini, a chance de um elemento escolhido aleatoriamente ser identificado incorretamente é de aproximadamente 69%

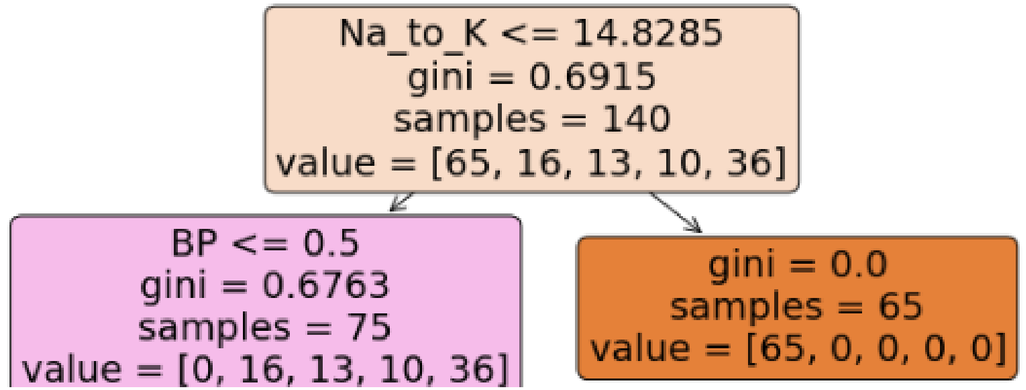
P(Y) = 65/140
P(A) = 16/140
P(B) = 13/140
P(C) = 10/140
P(X) = 36/140

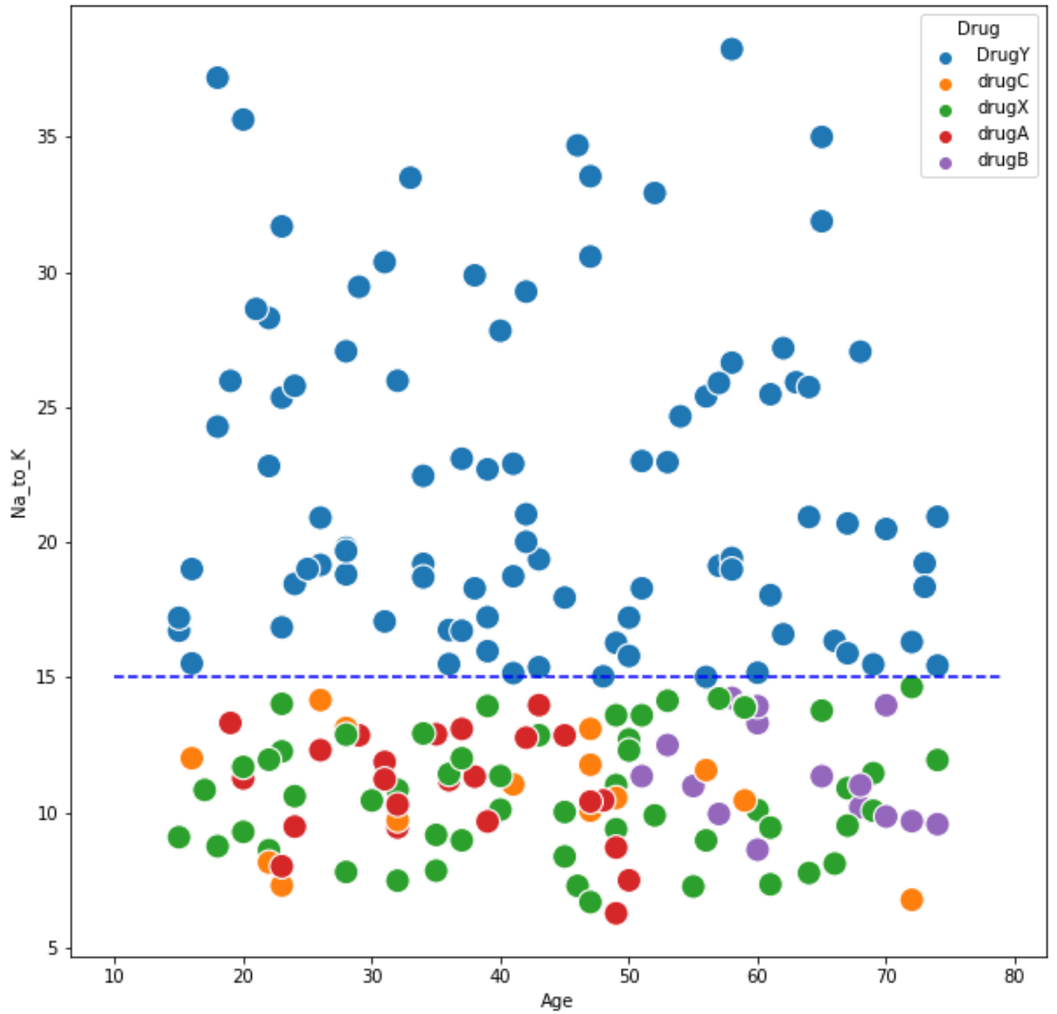
Gini = 1 - (65/140)^2 - (16/140)^2 - (13/140)^2 - (10/140)^2 - (36/140)^2 =

0,6915

Decisão da árvore

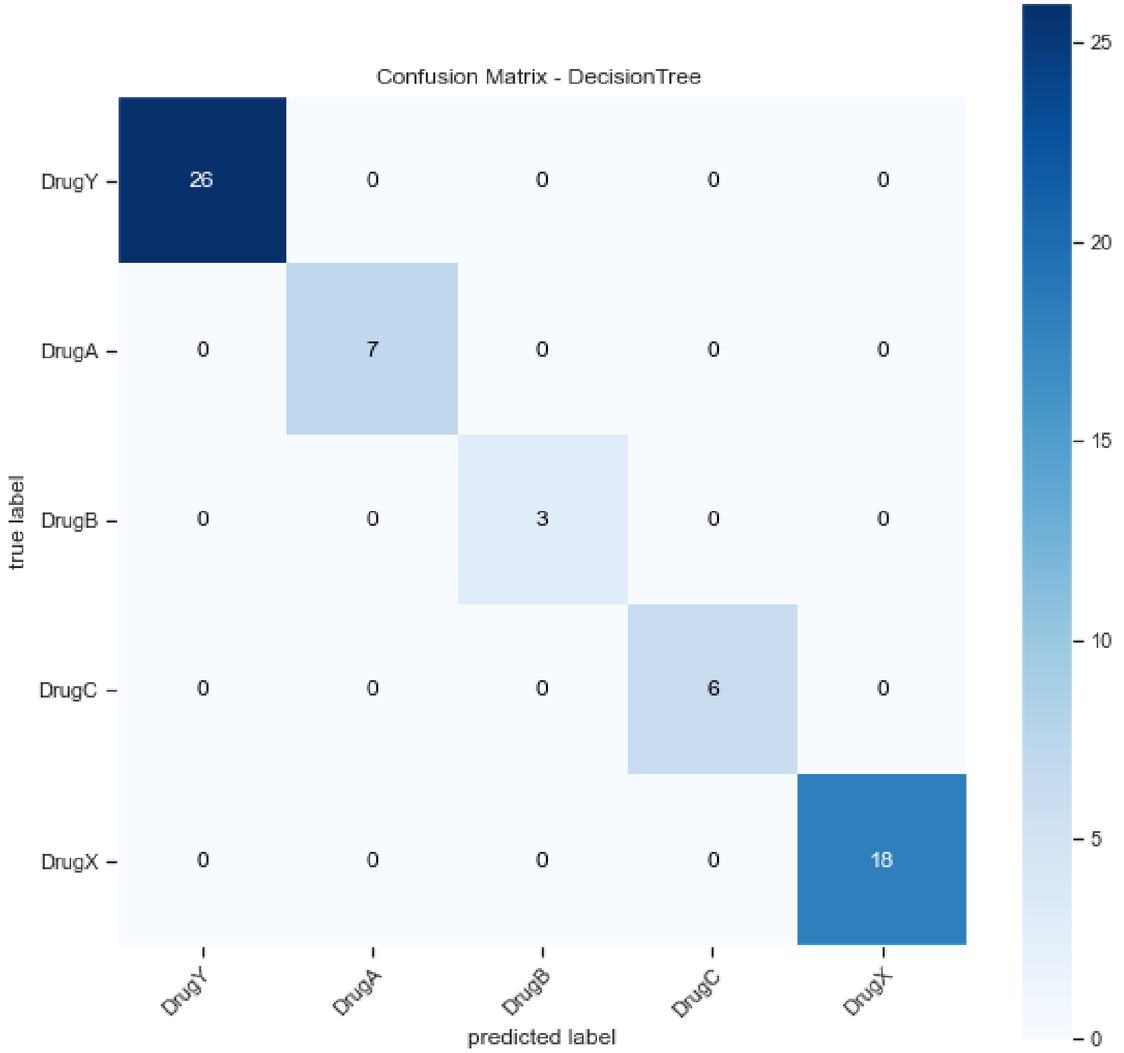
- Divisão perfeita entre a classe y e as demais classes





Matriz de Confusão

- **100% de Acurácia**



$$acurácia = \frac{Total\ de\ acertos}{Total\ de\ itens}$$

Outras Métricas

- **Precisão 100%**
- **Sensibilidade 100%**
- **F-score 100%**
- Obs. pré-processamento adicional não foi testado, pois o modelo já obteve máximo em todas as métricas;

	precision	recall	f1-score	support
0	1.00	1.00	1.00	26
1	1.00	1.00	1.00	7
2	1.00	1.00	1.00	3
3	1.00	1.00	1.00	6
4	1.00	1.00	1.00	18
accuracy			1.00	60
macro avg	1.00	1.00	1.00	60
weighted avg	1.00	1.00	1.00	60

Naive Bayes Gaussiano

- Assume que os atributos possuem distribuição Gaussiana
- Treinamento, teste e avaliação
- Cálculo de métricas de avaliação
- Análise visual dos resultadoas com matriz de confusão

Métricas sem Pré-Processamento Adicional

- **Menor precisão, falsos positivos;**
- **Maior recall, falsos negativos;**
- **Acurácia de 90%, possível overfitting;**

Gaussian Naives Bayes

	precision	recall	f1-score	support
0	1.00	0.77	0.87	26
1	0.88	1.00	0.93	7
2	0.50	1.00	0.67	3
3	0.75	1.00	0.86	6
4	1.00	1.00	1.00	18
accuracy			0.90	60
macro avg	0.82	0.95	0.87	60
weighted avg	0.94	0.90	0.90	60

Métricas com Pré-Processamento Adicional

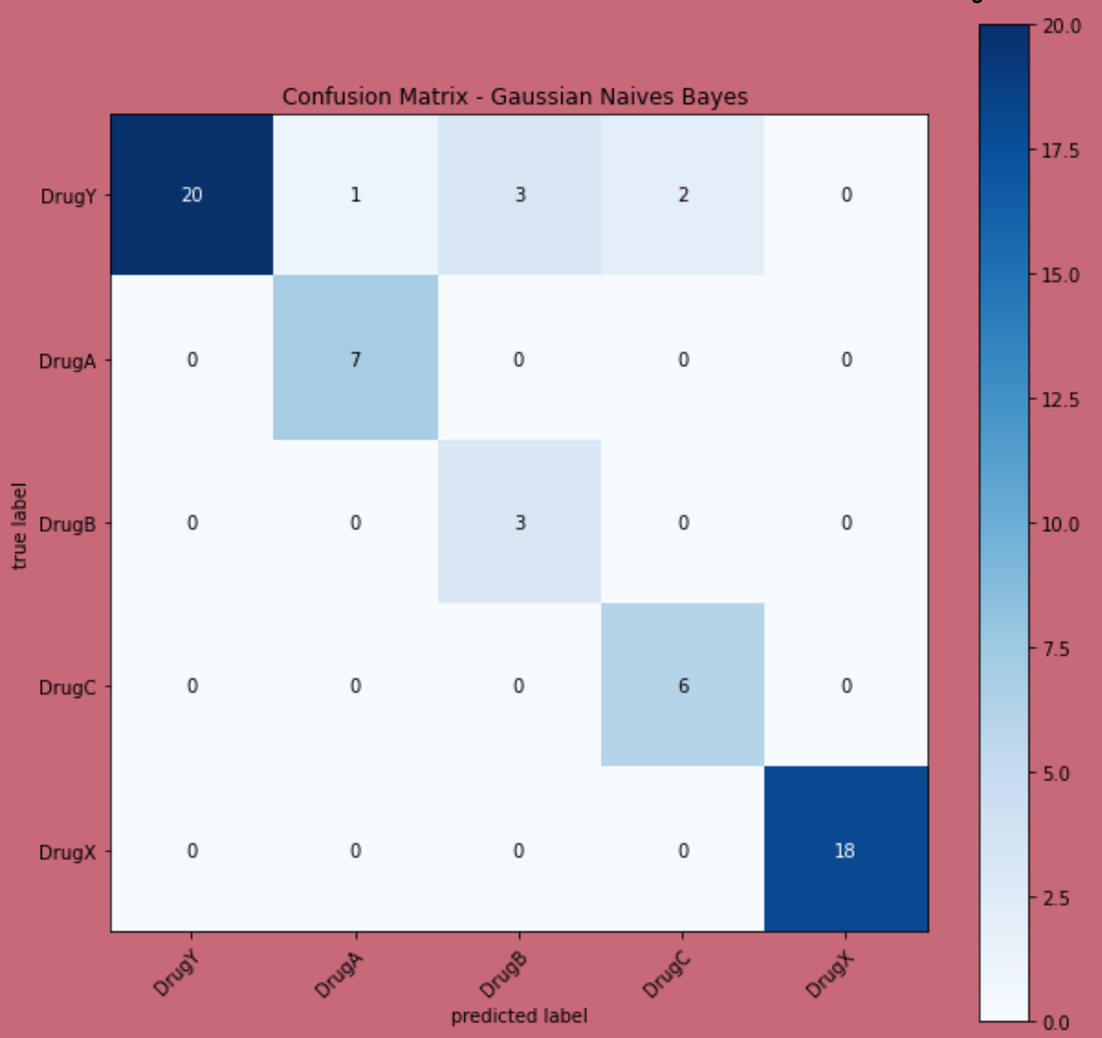
- Aumenta da precisão, menos falsos positivos;
- Queda do recall, mais falsos negativos;
- Aumento da acurácia, efeito do pré-processamento adicional;
- Acurácia de 92% atenuação de possível overfitting;

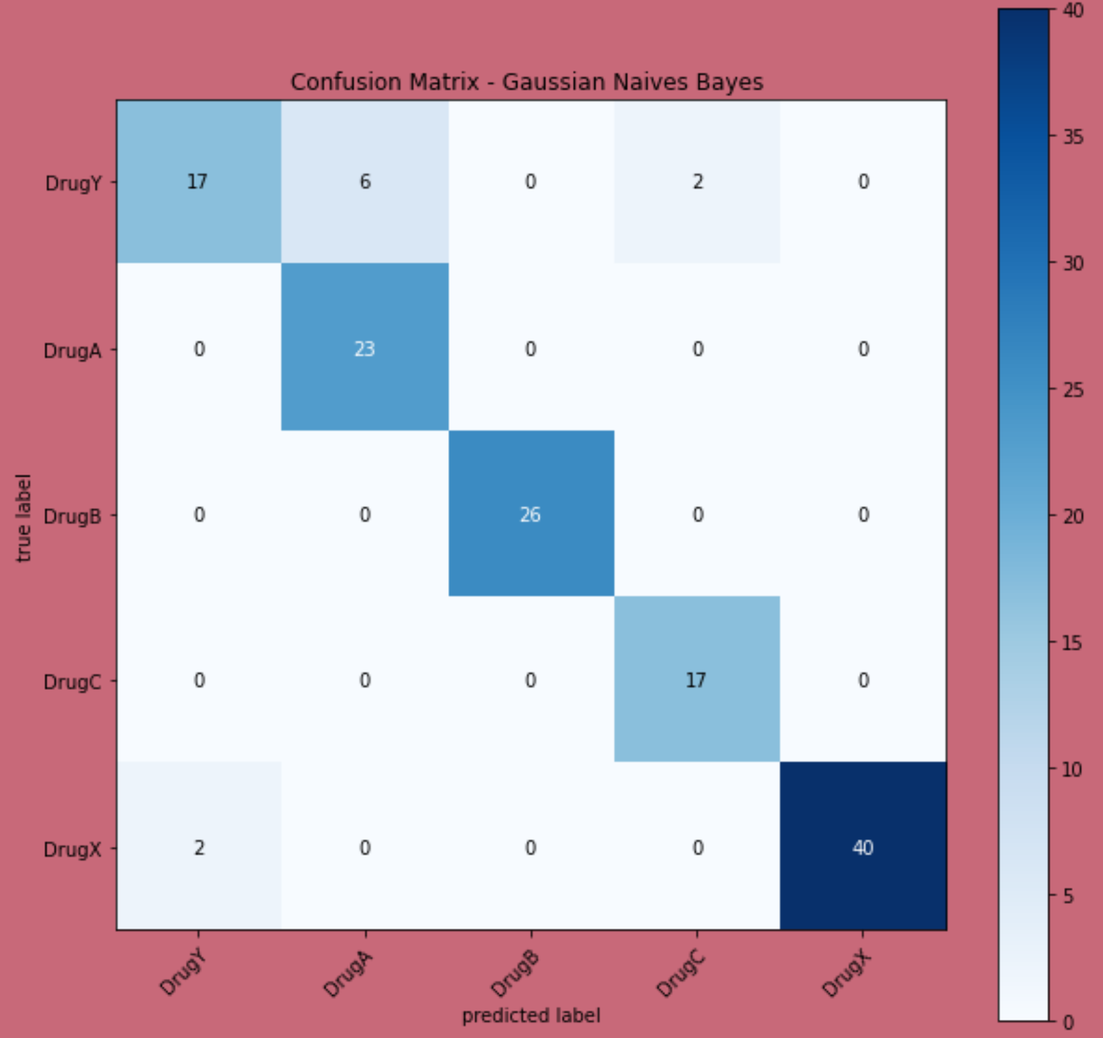
Gaussian Naives Bayes with additional pre-process

	precision	recall	f1-score	support
0	0.89	0.68	0.77	25
1	0.79	1.00	0.88	23
2	1.00	1.00	1.00	26
3	0.89	1.00	0.94	17
4	1.00	0.95	0.98	42
accuracy			0.92	133
macro avg	0.92	0.93	0.92	133
weighted avg	0.93	0.92	0.92	133

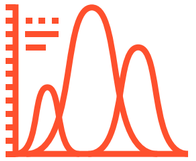
Sem pré-proc. adi. x Com pré-proc. adi.

Avalição com matriz de confusão





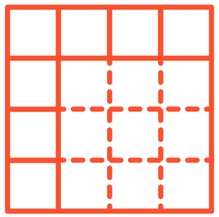
K Vizinhos Mais Próximos



Alguma particularidade do modelo

Treinamento, teste e avaliação

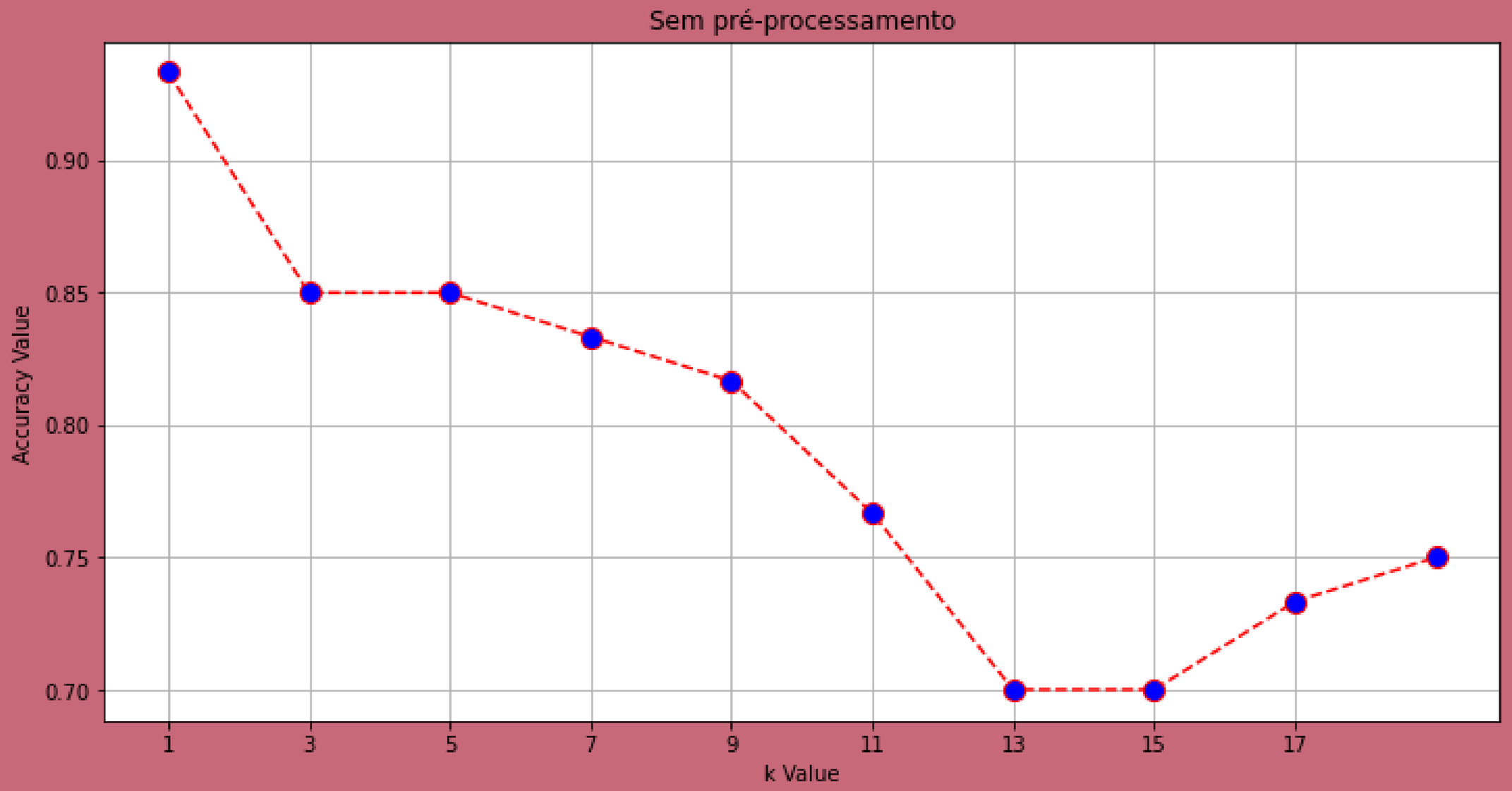
Cálculo de métricas de avaliação



Análise visual dos resultadoas com matriz de confusão

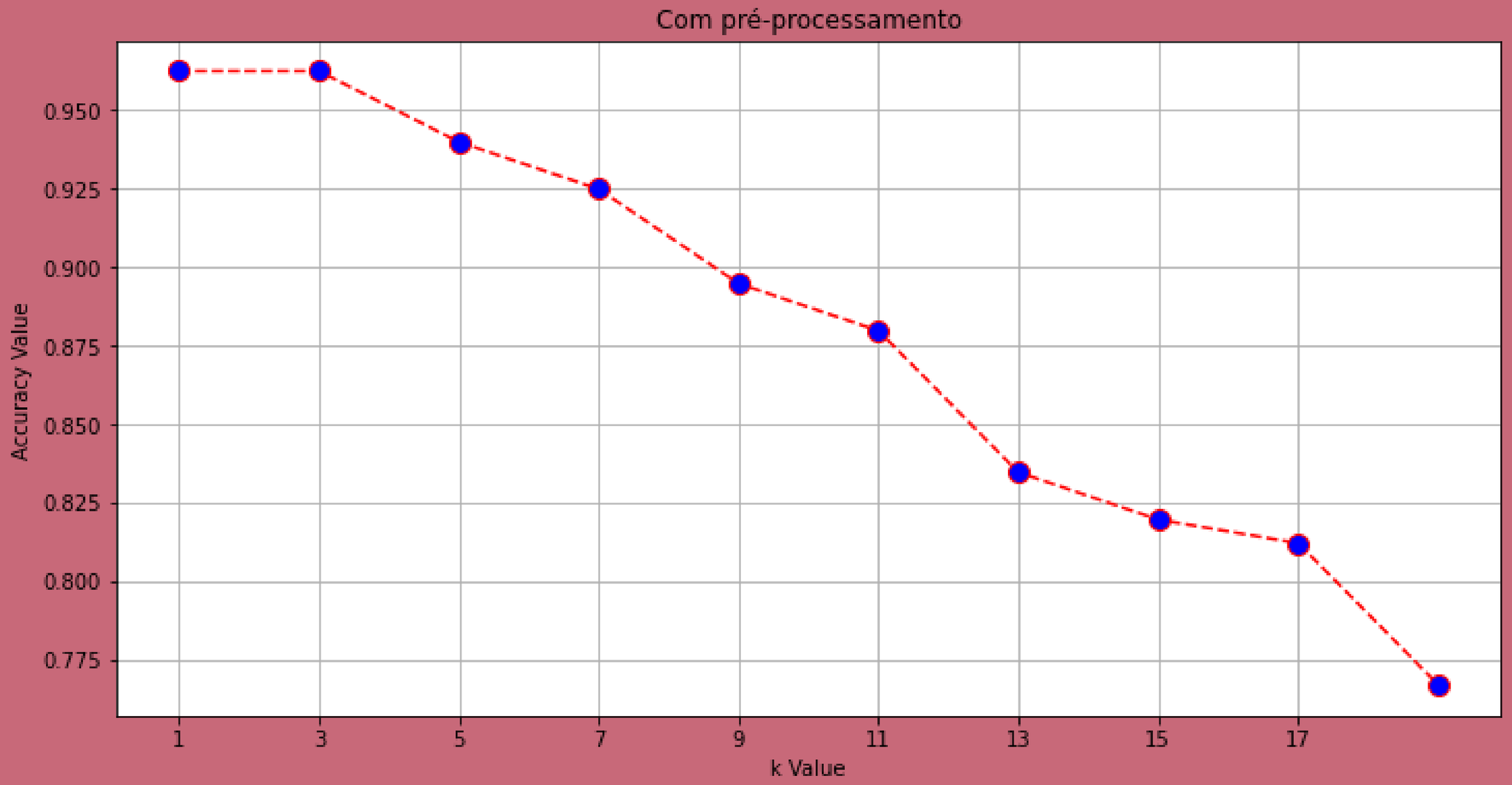
Sem pré-proc. adi. x Com pré-proc. adi.

Seleção da quantidade k de vizinhos



Sem pré-proc. adi. x Com pré-proc. adi.

Seleção da quantidade k de vizinhos



Métricas sem Pré-Processamento Adicional

- **k = 5**
- **Falsos positivos e negativos semelhantes em ambos os casos;**
- **Acurácia de 85%**

Sem Pré-processamento adicional

	precision	recall	f1-score	support
0	0.88	0.81	0.84	26
1	0.88	1.00	0.93	7
2	0.60	1.00	0.75	3
3	1.00	0.50	0.67	6
4	0.85	0.94	0.89	18
accuracy			0.85	60
macro avg	0.84	0.85	0.82	60
weighted avg	0.87	0.85	0.85	60

Métricas com Pré-Processamento Adicional

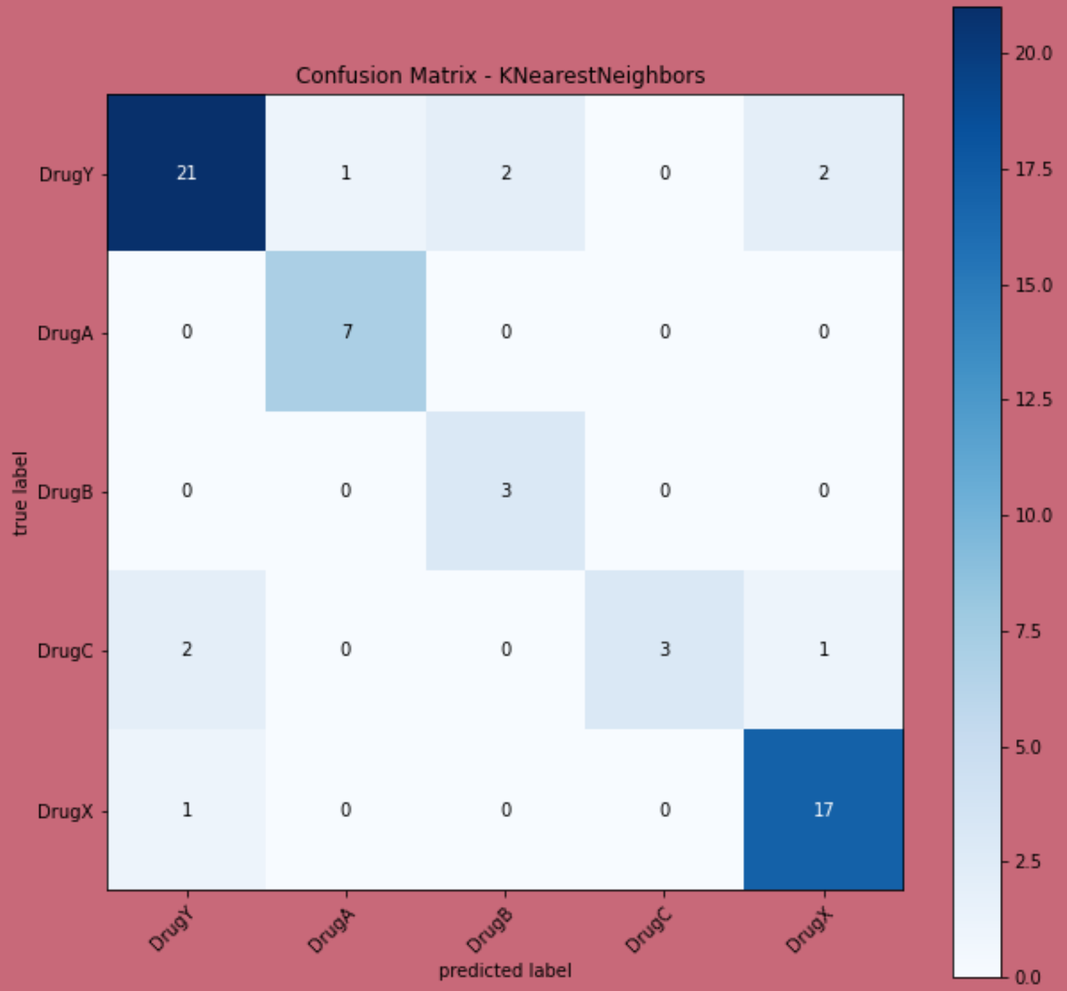
- **k = 9**
- **Aumento da acurácia, efeito do pré-processamento adicional;**
- **Acurácia de 89%;**

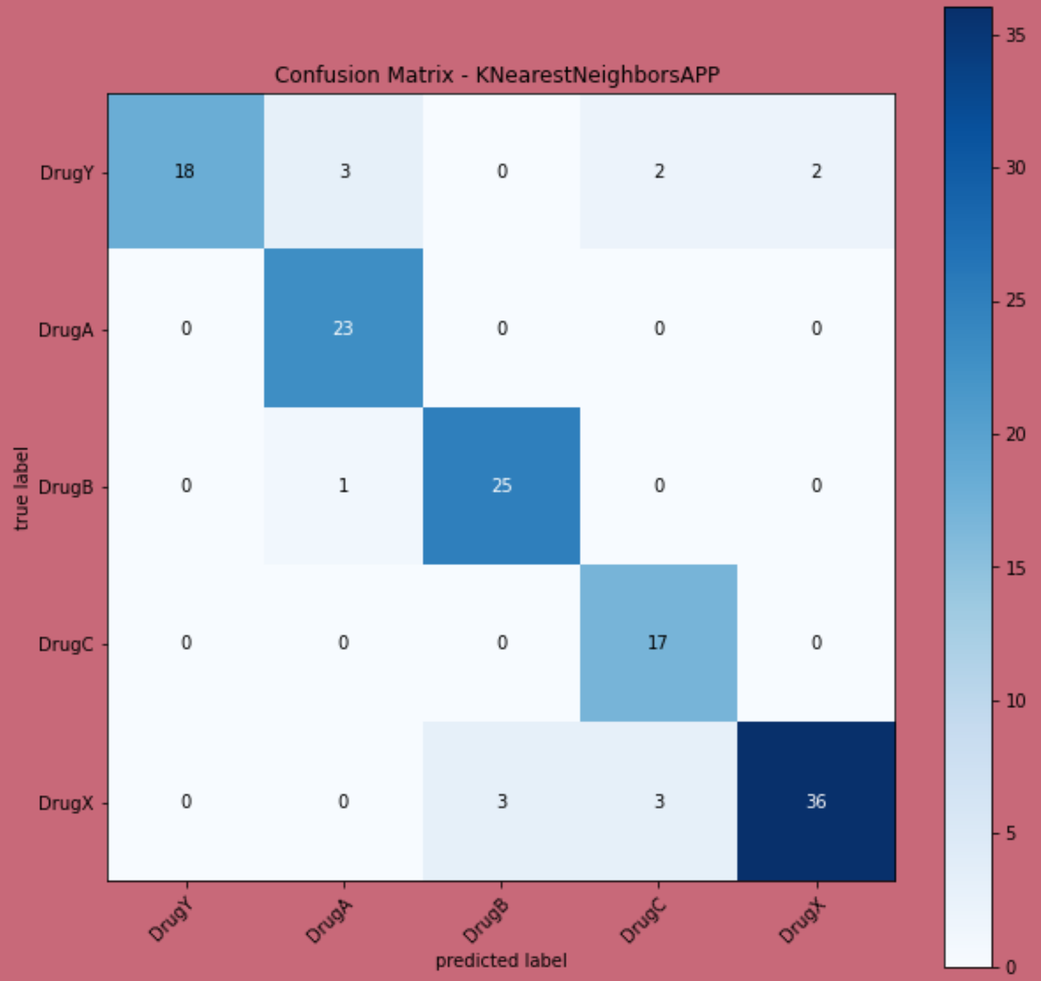
Com Pré-processamento adicional

	precision	recall	f1-score	support
0	1.00	0.72	0.84	25
1	0.85	1.00	0.92	23
2	0.89	0.96	0.93	26
3	0.77	1.00	0.87	17
4	0.95	0.86	0.90	42
accuracy			0.89	133
macro avg	0.89	0.91	0.89	133
weighted avg	0.91	0.89	0.89	133

Sem pré-proc. adi. x Com pré-proc. adi.

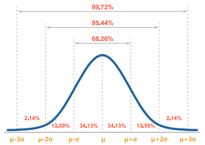
Avalição com matriz de confusão



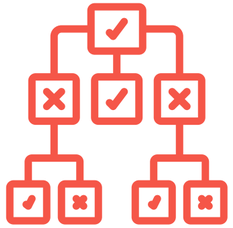


Considerações Finais

KNN pode ser bastante preciso ou ser capaz de generalizar, mostrou ser bastante adaptável a qualquer demanda;



GNB bastante preciso, pouca capacidade de generalização;



AD muito apto a dividir os dados, extremamente ajustado ao conjunto de dados;

Opinião de especialista para escolher o melhor modelo dado o contexto e possíveis novos dados;