Aprendizado de Maquina para Qualidade de Vinhos

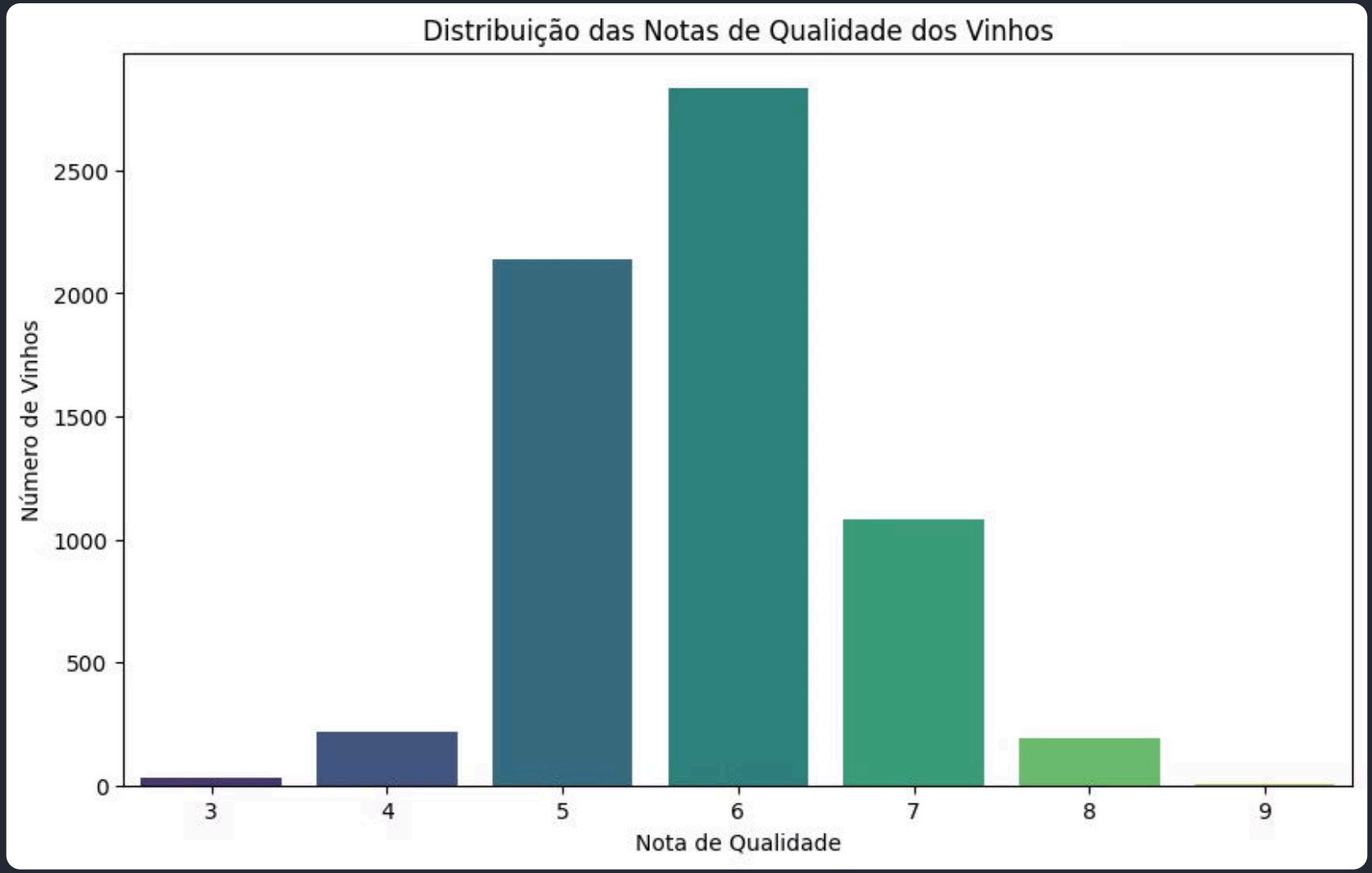
Arthur Bezerra Calado

Gabriel D'assumpção de Carvalho

Pedro Henrique Sarmento de Paula

Nome da variável	Papel	Tipo	Descrição	Valores ausentes
acidez fixa	Característica	Contínuo		Não
acidez volátil	Característica	Contínuo		Não
ácido cítrico	Característica	Contínuo		Não
açúcar residual	Característica	Contínuo		Não
Cloretos	Característica	Contínuo		Não
Dióxido de enxofre livre	Característica	Contínuo		Não
Dióxido de enxofre total	Característica	Contínuo		Não
densidade	Característica	Contínuo		Não
ph	Característica	Contínuo		Não
Sulfatos	Característica	Contínuo		Não
álcool	Característica	Contínuo		Não
qualidade	Alvo	Categórico	escore entre 0 e 10	Não
Cor	Característica	Categórico	vermelho (1) ou branco (0)	Não

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	color
count	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000	6497.0000
mean	7.2153	0.3397	0.3186	5.4432	0.0560	30.5253	115.7446	0.9947	3.2185	0.5313	10.4918	0.2461
std	1.2964	0.1646	0.1453	4.7578	0.0350	17.7494	56.5219	0.0030	0.1608	0.1488	1.1927	0.4308
min	3.8000	0.0800	0.0000	0.6000	0.0090	1.0000	6.0000	0.9871	2.7200	0.2200	8.0000	0.0000
25%	6.4000	0.2300	0.2500	1.8000	0.0380	17.0000	77.0000	0.9923	3.1100	0.4300	9.5000	0.0000
50%	7.0000	0.2900	0.3100	3.0000	0.0470	29.0000	118.0000	0.9949	3.2100	0.5100	10.3000	0.0000
75%	7.7000	0.4000	0.3900	8.1000	0.0650	41.0000	156.0000	0.9970	3.3200	0.6000	11.3000	0.0000
max	15.9000	1.5800	1.6600	65.8000	0.6110	289.0000	440.0000	1.0390	4.0100	2.0000	14.9000	1.0000



Agrupamento de Classes

Notas Baixas

Classe 0: [0, 4]

Notas Médias

Classe 1: [5, 7]

Notas Altas

Classe 2: [8, 10]

O objetivo é simplificar a previsão da qualidade do vinho agrupando as notas em categorias.

Árvores de Decisão

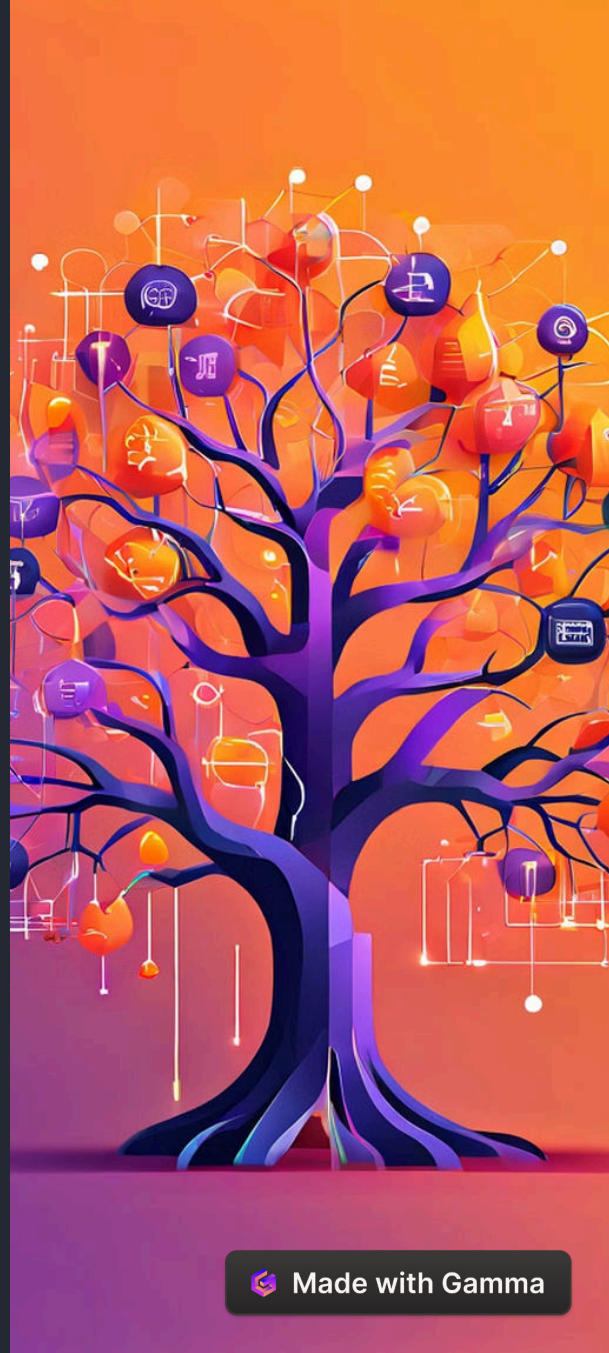
Metodologia

Separação dos dados em treino (70%) e teste (30%). Teste com diferentes valores de max_depth (10, 20, 40, 80, 160) e critérios (gini e entropy)

Treinamento

Usado o modelo DecisionTreeClassifier.

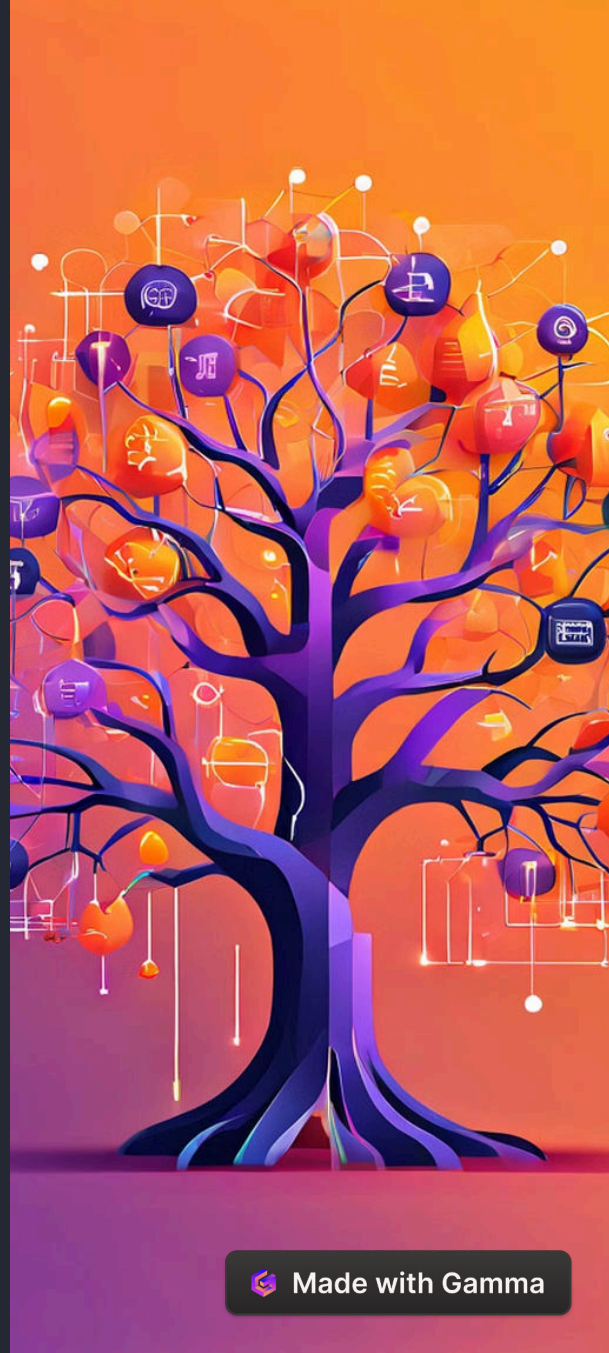
Ajuste de parâmetros como profundidade máxima e critérios de divisão para avaliar o impacto na precisão e cobertura.



Árvores de Decisão

Avaliação

- Resultados para o critério gini:
  - max_depth = 10: Precisão de 88.78%, recall de 91.49%.
  - max_depth = 20: Precisão de 90.79%, recall de 90.41%.
  - max_depth >= 40: Precisão estabilizada em 91.23%, recall decrescendo indicando possíveis sinais de overfitting à medida que a árvore se torna mais profunda.



Árvores de Decisão

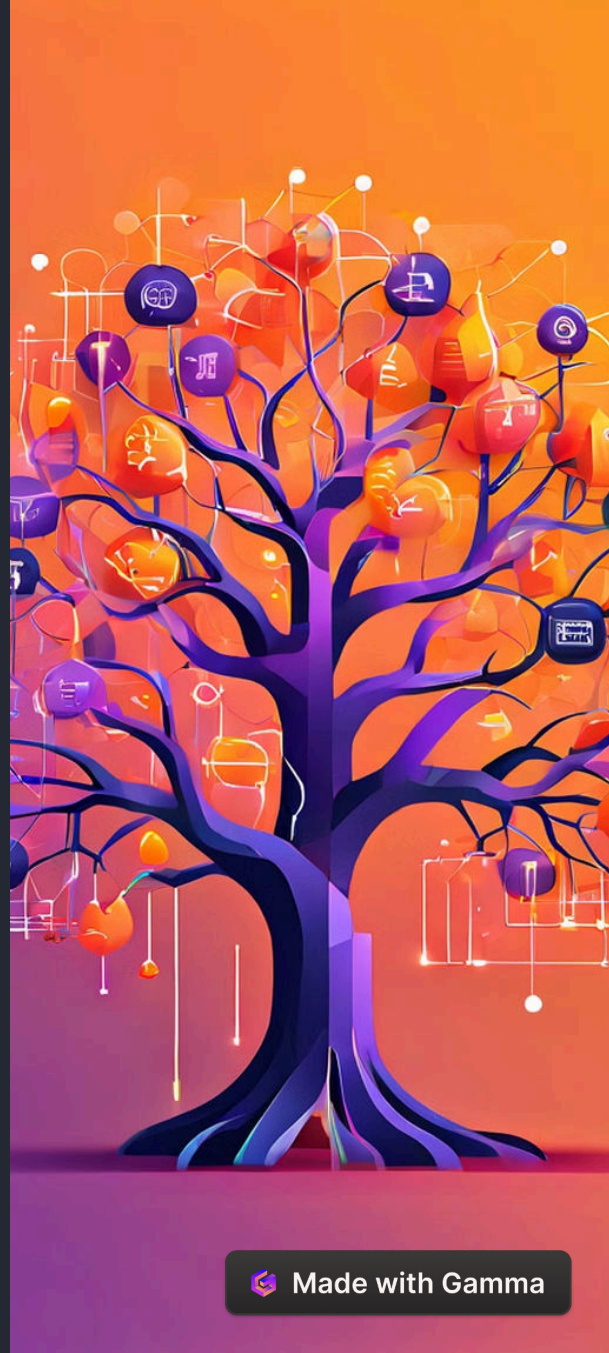
Avaliação

- Resultados para o critério entropy:
  - max_depth = 10: Precisão de 89.52%, recall de 91.44%.
  - max_depth = 20, 40, 80, 160: A precisão permanece relativamente estável em torno de 90.53%, independentemente da profundidade máxima da árvore. O recall diminui para cerca de 89.64% à medida que a profundidade aumenta.

Árvores de Decisão

Conclusões

- Árvores mais profundas podem levar a overfitting.
- Critério gini teve melhor desempenho geral que entropy.
- Gini: max_depth: 22; precision: 91.30%; recall: 90.56%;
- Entropy: max_depth: 8; precision: 90.25%; recall: 92.56%;



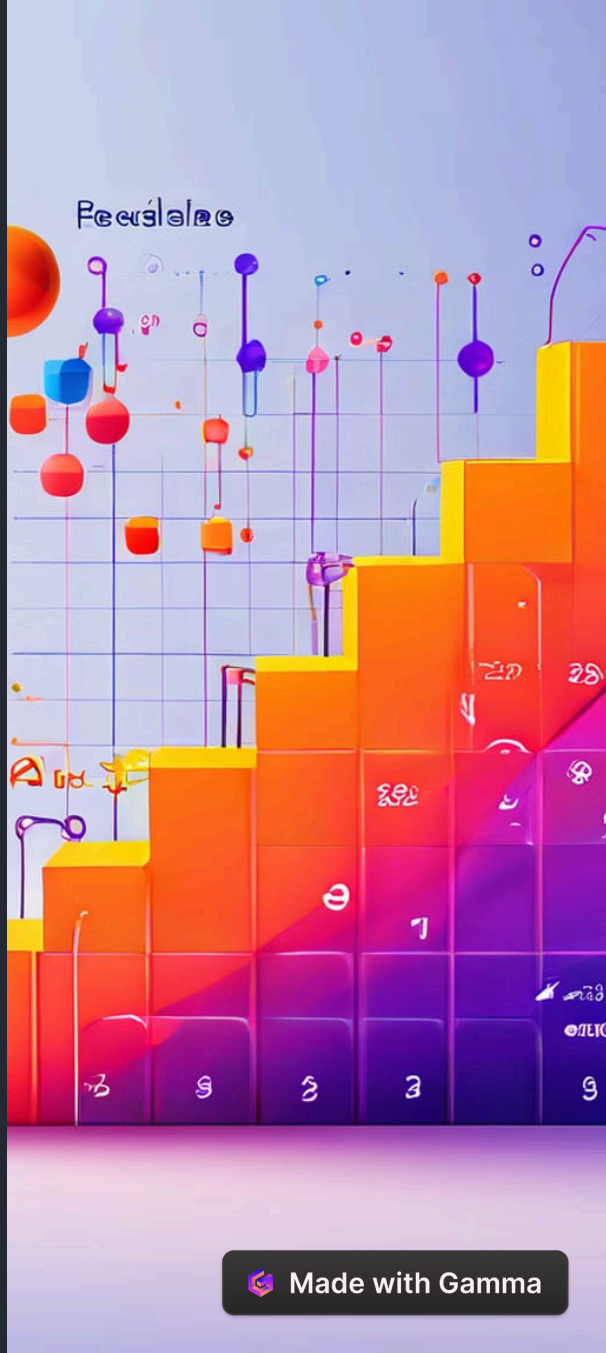
Bayesiano Ingênuo experimento 1

Metodologia e treinamento

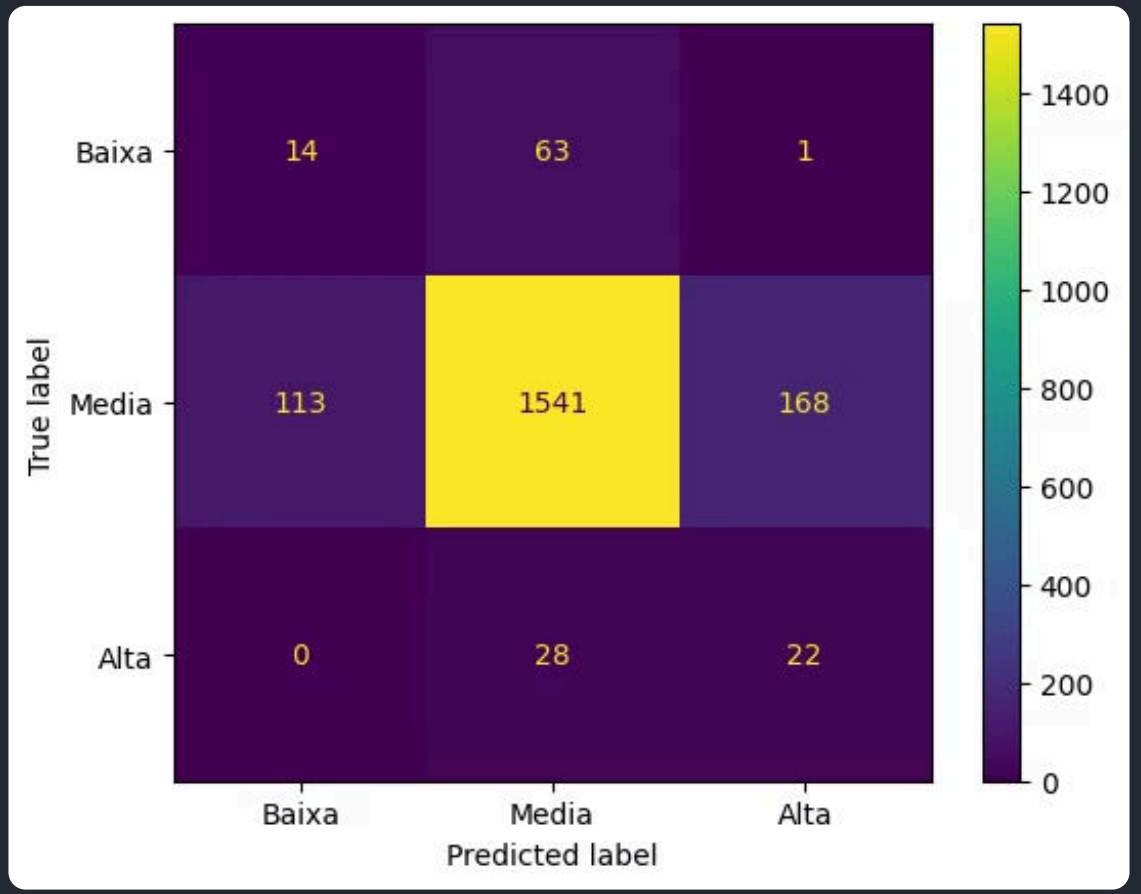
Divisão dos dados em conjuntos de treino e teste (70-30)

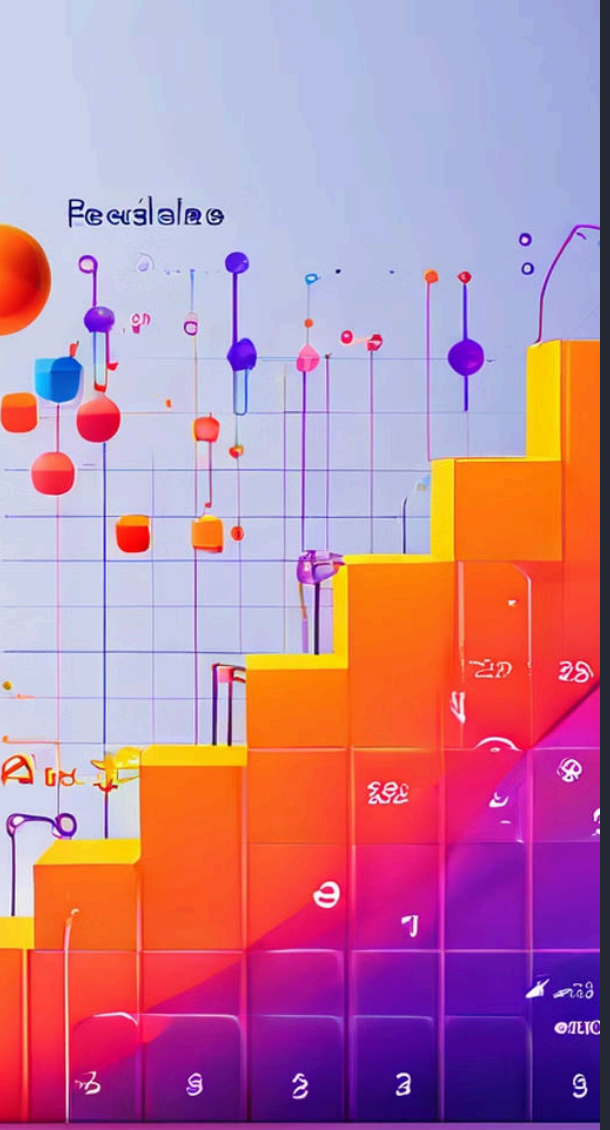
Função para fornecer uma análise do comportamento do modelo (precisão, cobertura e F1-Score)

Naive Bayes Gaussiano foi escolhido para este experimento devido à sua capacidade de lidar com uma combinação de variáveis contínuas e binárias, como as presentes neste conjunto de dados



Bayesiano Ingêuo experimento 1



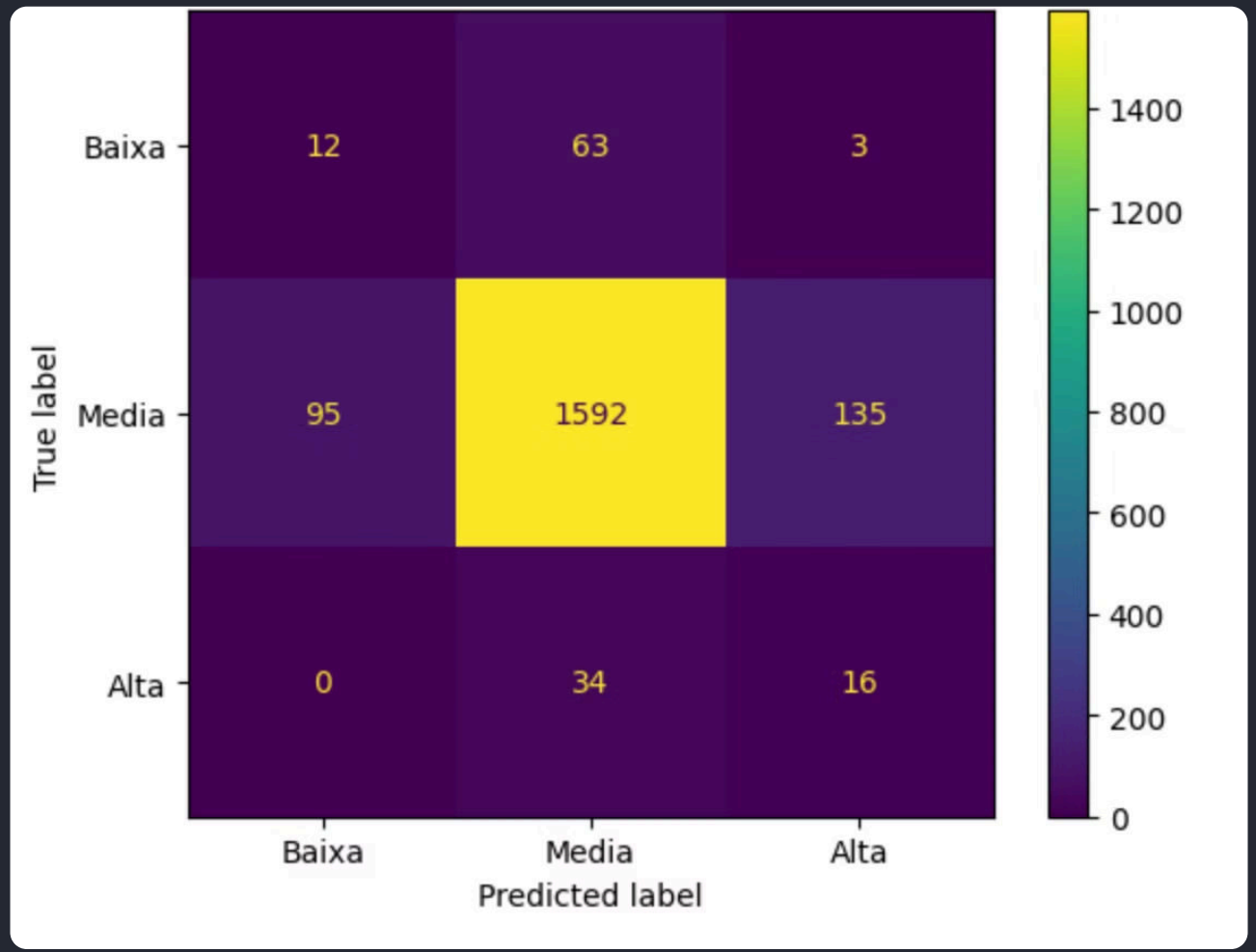


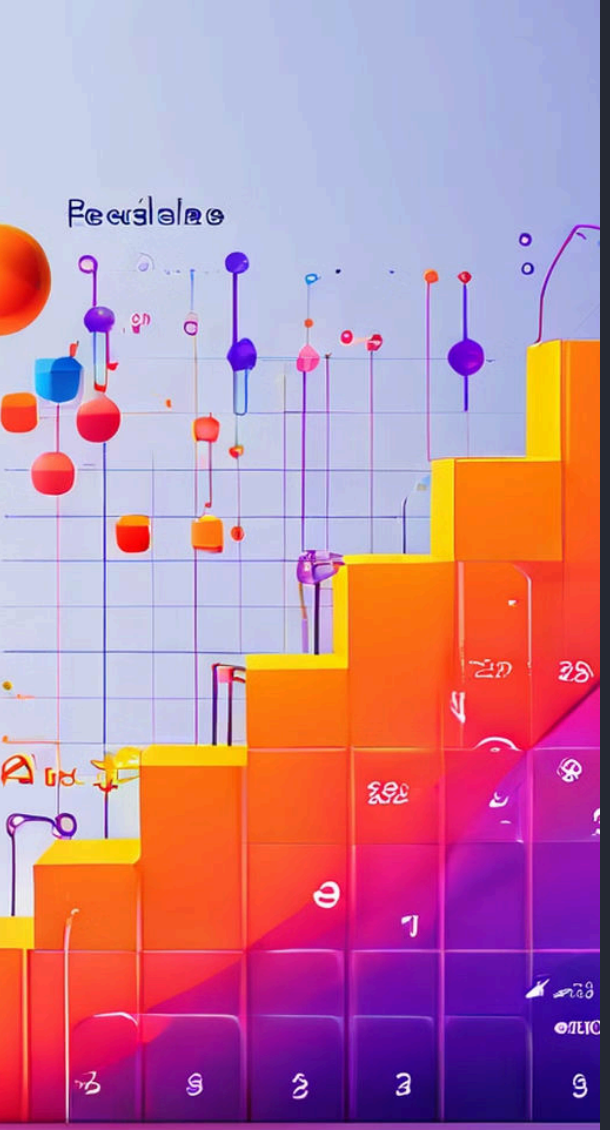
Bayesiano Ingênuo experimento 1

Resultados

Os resultados do primeiro experimento com o modelo mostram uma precisão de 88,96%, recall de 80,87% e F1-Score de 84,39%.

Bayesiano Ingêuo experimento 2





Bayesiano Ingênuo experimento 2

Resultados

As transformações nos dados no segundo experimento resultaram em melhorias significativas nas métricas do modelo. A precisão manteve-se alta, passando de 88,96% para 88,78%, enquanto o recall aumentou de 80,87% para 83,08%, indicando que o modelo identificou mais instâncias positivas. Consequentemente, o F1-Score melhorou de 84,39% para 85,65%, refletindo um melhor equilíbrio entre precisão e recall.

Bayesiano Ingênuo experimento 3

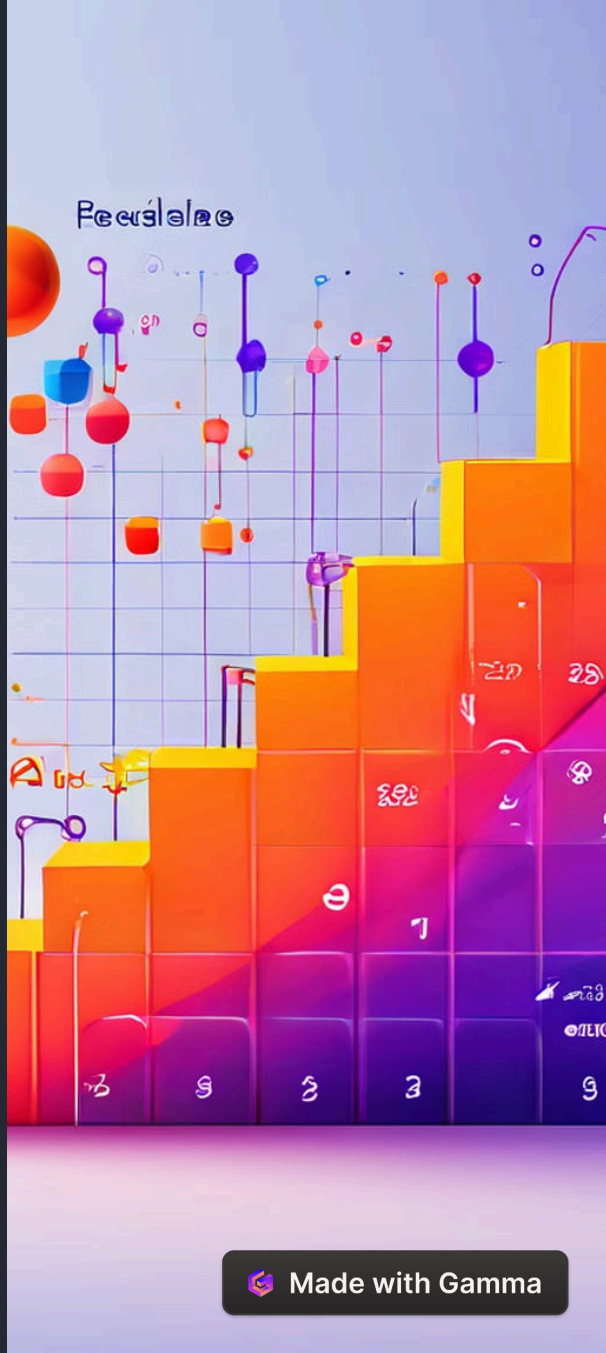
Metodologia e treinamento

Divisão dos dados em conjuntos de treino e teste (70-30)

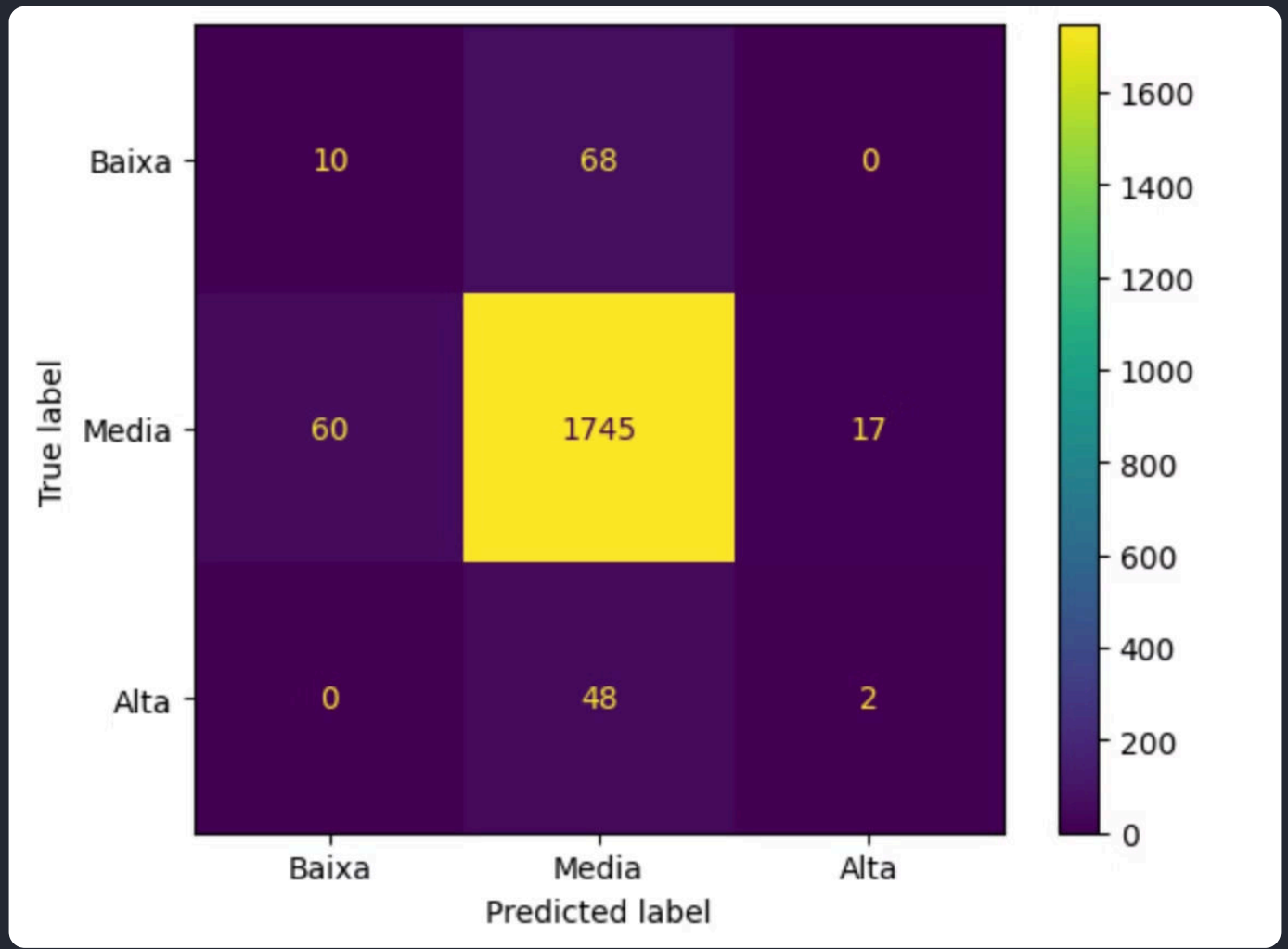
Função para fornecer uma análise do comportamento do modelo (precisão, cobertura e F1-Score)

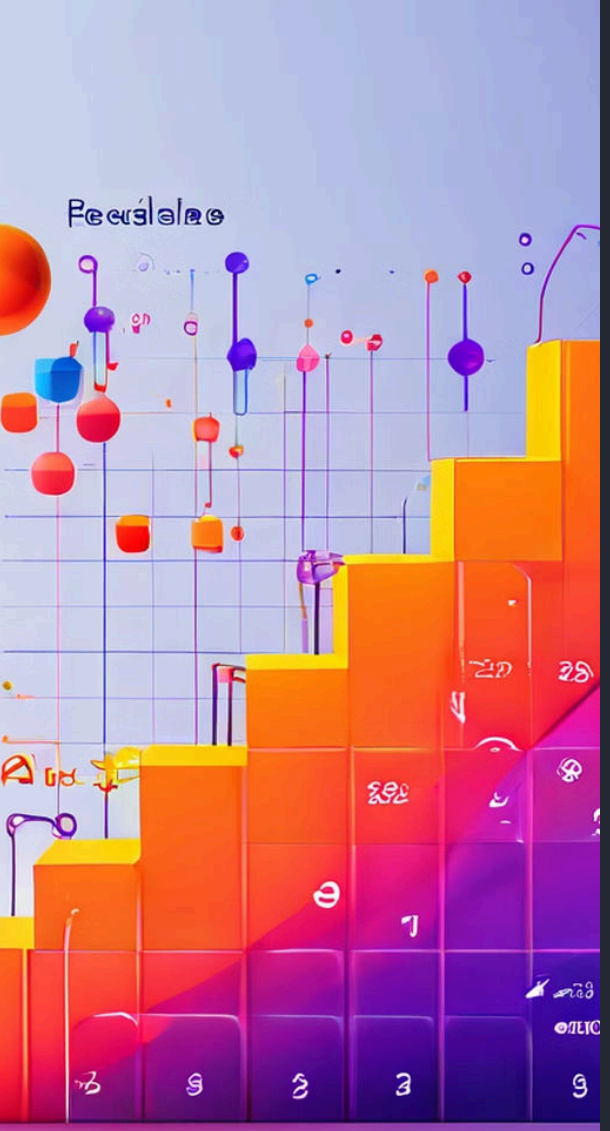
SelectKBest com o critério f_classif para mostrar as pontuações de relevância de cada informação nos dados para a predição

5 características com maior relevância, baseado na influência na variável-alvo, focando nas características mais informativas



Bayesiano Ingêuo experimento 3





Bayesiano Ingênuo experimento 3

Resultados

No resultado final do experimento com o Modelo Bayesiano Ingênuo, obtivemos uma precisão de 88,45%, recall de 90,10% e F1-Score de 89,23%.

Regressão Logística 1

Metodologia

Aplicação de regressão logística para prever as classes de qualidade de vinho. Utilização da função sigmoide para transformar a saída linear em probabilidade de classe. Foram usados dados sem transformação e com transformação para comparação. Empregou-se a técnica de validação cruzada

Regressão Logística 1

Treinamento

- Divisão dos dados em conjuntos de treino e teste (80-20).
- Ajuste dos parâmetros do modelo.
- Treinamento do modelo com o conjunto de treino utilizando a função de custo log-likelihood.
- Ajuste dos parâmetros do modelo via métodos de otimização como gradiente descendente.
- Os resultados mostram que o modelo se ajustou bem aos dados.
- Score com base em validação cruzada (5 folds).

Regressão Logística 1

Avaliação

- Avaliação do modelo com base em precisão e recall.

Conclusões

- Desempenho aceitável em termos de precisão e recall.
- Regressão logística pode não capturar todas detalhes pela natureza linear do modelo.
- Os resultados mostram que o modelo se ajustou bem aos dados. Ele obteve uma precisão notável, alcançando aproximadamente 90,26% de acurácia com base na métrica F1. Além disso, durante a validação cruzada, o modelo manteve um desempenho consistente, com uma média de 93,05% nos diferentes agrupamentos.

```
Coeficientes para a classe 1:
fixed acidity: 0.2706
volatile acidity: 0.1678
citric acid: -0.0494
residual sugar: -0.0500
chlorides: 0.0080
free sulfur dioxide: -0.0418
total sulfur dioxide: 0.0090
density: 0.0175
pH: 0.0567
sulphates: -0.0485
alcohol: -0.2595
color: -0.0622

Coeficientes para a classe 2:
fixed acidity: 0.1721
volatile acidity: -0.0875
citric acid: 0.0291
residual sugar: -0.0011
chlorides: 0.0132
free sulfur dioxide: 0.0115
total sulfur dioxide: 0.0015
density: 0.0693
pH: 0.2443
sulphates: 0.1030
alcohol: -0.0355
color: 0.2448

Coeficientes para a classe 3:
fixed acidity: -0.4427
volatile acidity: -0.0803
citric acid: 0.0202
residual sugar: 0.0511
chlorides: -0.0212
free sulfur dioxide: 0.0302
total sulfur dioxide: -0.0105
density: -0.0868
pH: -0.3010
sulphates: -0.0545
alcohol: 0.2950
color: -0.1825
```

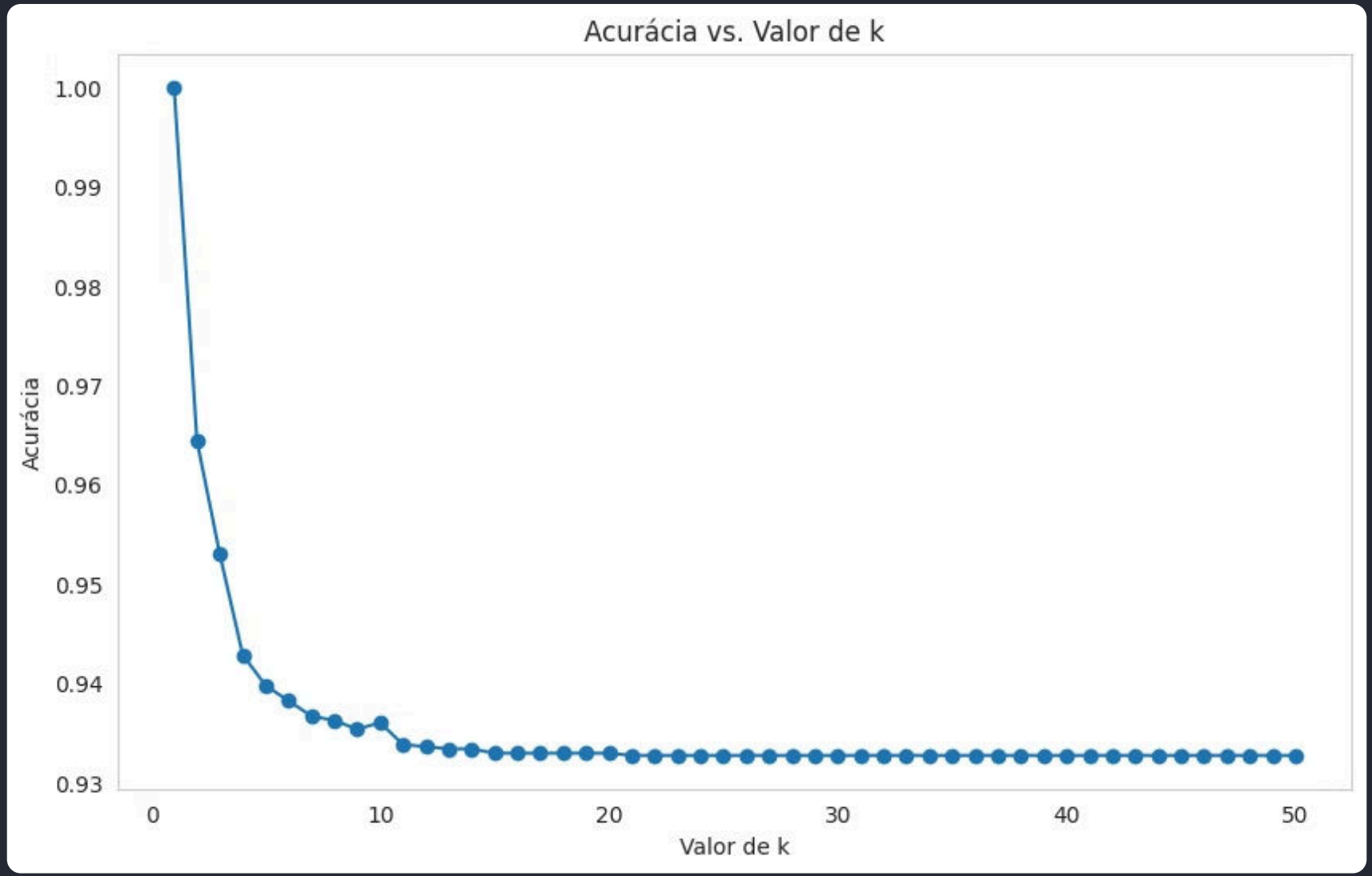
K Vizinhos

Metodologia

Uso do algoritmo K-Nearest Neighbors (KNN) para classificação, testando diferentes valores de k. Uso das distâncias euclidianas.

Treinamento

Divisão dos dados (treino 70% e teste 30%), padronização (StandardScaler) e ajuste do número de vizinhos k (1 a 50). K=1 foi resultou na maior **acurácia (100%)** no treinamento, podendo indicar overfitting.



K Vizinhos

Avaliação

Análise de precisão, recall e matriz de confusão para diferentes valores de k.

Resultados

- Modelo sensível à escolha de, escala dos dados, e ruídos nos dados.
- Usando o melhor k (k=1), o modelo foi avaliado no conjunto de teste. Métricas de desempenho:
  - **Acurácia: 91.18%**; Precisão: 90.65%
  - Recall: 91.18%; F1 Score: 90.88%