

# Algoritmos Supervisionados – Naive Bayes



Plataforma completa de aprendizado  
contínuo em programação.

**#BoostingPeople**

[rocketseat.com.br](https://rocketseat.com.br)

Todos os direitos reservados © Rocketseat S.A.

# Algoritmos Supervisionados

## Naive Bayes

O objetivo deste módulo é apresentar o algoritmo de **Naive Bayes**, baseado no conceito de probabilidade condicional e independência de variáveis, onde faremos o **processo completo**, desde o EDA, seleção automática de features até a entrega do modelo através de uma API.



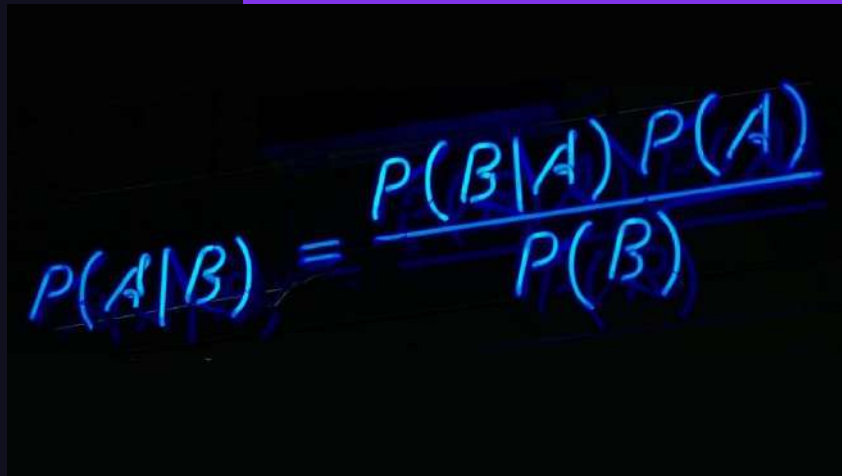
# Agenda

- O que é o algoritmo Naive Bayes?
- Métricas de Classificação
- Projeto – Naive Bayes



# O que é o algoritmo de Naive Bayes?

O algoritmo **Naive Bayes** é um método de aprendizado de máquina que faz previsões com base no **Teorema de Bayes**. Ele assume que as características (variáveis) usadas para fazer a previsão **são independentes entre si**, o que é uma simplificação (daí o termo "naive" ou ingênuo). O modelo calcula a probabilidade de uma instância pertencer a uma classe específica usando a **probabilidade condicional** das características dadas essa classe. Naive Bayes é amplamente utilizado em classificação de textos, como filtragem de spam, e em outras aplicações onde a independência simplificada é aceitável. É eficiente e fácil de implementar, **especialmente para conjuntos de dados grandes**.

A photograph of a piece of lined paper with the Naive Bayes formula written in blue ink. The formula is  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ . The paper is slightly tilted and has some faint, illegible markings in the background.
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

# O que é o algoritmo de Naive Bayes?

Probability **B** Will Happen Given  
Evidence **A** Has Already Happened      Probability **A** Will Happen

$$\underbrace{P(A|B)} = \frac{\overbrace{P(B|A)} \cdot \overbrace{P(A)}}{\underbrace{P(B)}}$$

Probability **A** Will Happen Given  
Evidence **B** Has Already Happened      Probability **B** Will Happen

© howstuffworks

# O que é o algoritmo de Naive Bayes?

Suponha que  $A$  seja a presença de uma doença e  $B$  seja um resultado positivo em um teste. As probabilidades podem ser interpretadas como:

- $P(A|B)$  é a probabilidade de ter a doença dado um resultado positivo no teste.
- $P(B|A)$  é a probabilidade de um resultado positivo no teste dado que a pessoa tem a doença.
- $P(A)$  é a probabilidade geral de ter a doença.
- $P(B)$  é a probabilidade geral de obter um resultado positivo no teste.

# O que é o algoritmo de Naive Bayes?

Suponha que  $A$  seja a presença de uma doença e  $B$  seja um resultado positivo em um teste. As probabilidades podem ser interpretadas como:

- $P(B|A) = 0.95$
- $P(A) = 0.01$
- $P(B) = 0.06$
- $P(A|B) = \frac{0.95 * 0.01}{0.06} = 0.1583$  ou 15,83%.

# O que é o algoritmo de Naive Bayes?

$$P(y \mid x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n \mid y)}{P(x_1, \dots, x_n)}$$



# Métricas de Classificação

## Acurácia

A acurácia é a proporção de instâncias corretamente classificadas em relação ao total de instâncias. Apesar de fácil interpretação, pode ser enganosa em conjuntos de dados desbalanceados, pois favorece a classe majoritária.

$$\frac{TP + TN}{Total\ de\ instâncias}$$

## Precisão

A precisão é a razão de instâncias verdadeiramente positivas em relação ao total de instâncias classificadas como positivas. A precisão é importante quando o custo de falsos positivos é alto.

$$\frac{TP}{TP + FP}$$

		P R E D I T O	
		👍 POSITIVO	👎 NEGATIVO
R E A L	👍 POSITIVO	✅ 👍 TP verdadeiro positivo	❌ 👎 FN falso negativo
	👎 NEGATIVO	❌ 👍 FP falso positivo	✅ 👎 TN verdadeiro negativo

# Métricas de Classificação

## Acurácia x Precisão

Um cenário em que a precisão é mais apropriada que a acurácia como métrica de validação ocorre quando há um desequilíbrio significativo nas classes do conjunto de dados. Suponha que você esteja desenvolvendo um modelo para detectar fraudes em transações financeiras, onde a maioria das transações é legítima (classe negativa) e apenas uma pequena porcentagem é fraudulenta (classe positiva).

Se o modelo classificar todas as transações como legítimas, a acurácia pode ser alta, pois a maioria das previsões estará correta. No entanto, isso não é útil, pois o objetivo é identificar as transações fraudulentas. Nesse caso, a precisão (número de transações fraudulentas corretamente identificadas dividido pelo total de transações identificadas como fraudulentas) seria uma métrica mais relevante, pois se concentra na performance da classe positiva.

Em resumo, quando há um desequilíbrio nas classes e o custo de falsos positivos é alto (no caso de fraudes, por exemplo), a precisão é uma métrica mais informativa do que a acurácia

		P R E D I T O	
		👍 POSITIVO	👎 NEGATIVO
R E A L	👍 POSITIVO	✅ 👍 TP verdadeiro positivo	❌ 👎 FN falso negativo
	👎 NEGATIVO	❌ 👍 FP falso positivo	✅ 👎 TN verdadeiro negativo

# Métricas de Classificação

## Recall

O Recall mede a proporção de instâncias verdadeiramente positivas que foram corretamente identificadas pelo modelo. É crucial quando o custo de falsos negativos é alto.

$$\frac{TP}{TP + FN}$$

## F1-Score

O F1-Score é a média harmônica entre precisão e revocação, fornecendo uma única medida que equilibra ambas. É especialmente útil quando há um desequilíbrio entre as classes.

$$2 \times \frac{\text{Precisão} * \text{Recall}}{\text{Precisão} + \text{Recall}}$$

		P R E D I T O	
		👍 POSITIVO	👎 NEGATIVO
R E A L	👍 POSITIVO	✅ 👍 TP verdadeiro positivo	❌ 👎 FN falso negativo
	👎 NEGATIVO	❌ 👍 FP falso positivo	✅ 👎 TN verdadeiro negativo

# Métricas de Classificação

## Acurácia x Precisão x Recall

O recall é uma métrica apropriada quando o custo associado aos falsos negativos é alto e o objetivo é minimizar a quantidade de casos positivos que são erroneamente classificados como negativos. Em cenários críticos, como na detecção de doenças graves ou na prevenção de fraudes, priorizar o recall é fundamental.

Por exemplo, considere um modelo para detectar uma doença rara. Se o modelo erradamente classificar um paciente como saudável quando ele realmente está doente (falso negativo), as consequências podem ser graves. Nesse caso, maximizar o recall é crucial para garantir que a maioria dos casos positivos seja corretamente identificada, mesmo que isso resulte em alguns falsos positivos.

Em resumo, escolha recall quando o foco está em identificar corretamente a maioria dos casos positivos, especialmente quando os falsos negativos têm consequências significativas. A acurácia pode ser enganosa em situações desbalanceadas, e a precisão pode ser mais apropriada quando o custo de falsos positivos é elevado.

		P R E D I T O	
		👍 POSITIVO	👎 NEGATIVO
R E A L	👍 POSITIVO	✅ 👍 TP verdadeiro positivo	❌ 👎 FN falso negativo
	👎 NEGATIVO	❌ 👍 FP falso positivo	✅ 👎 TN verdadeiro negativo

# Métricas de Classificação

## Acurácia x Precisão x Recall x F1-Score

O F1-Score é uma métrica apropriada quando há um equilíbrio desejado entre precisão e recall, e ambas são igualmente importantes. Um exemplo é a classificação de modelos de reconhecimento de spam em e-mails.

Nesse cenário, é crucial alcançar um equilíbrio entre identificar a maioria dos e-mails de spam (recall) e garantir que os e-mails identificados como spam sejam, de fato, spam (precisão). A acurácia pode ser enganosa, especialmente se a maioria dos e-mails não for spam. O F1-Score considera tanto a precisão quanto o recall, proporcionando uma métrica única que equilibra essas duas considerações.

Em resumo, o F1-Score é a escolha adequada quando há um compromisso necessário entre precisão e recall, e ambos são cruciais para a avaliação do desempenho do modelo.

		P R E D I T O	
		👍 POSITIVO	👎 NEGATIVO
R E A L	👍 POSITIVO	✅ 👍 TP verdadeiro positivo	❌ 👎 FN falso negativo
	👎 NEGATIVO	❌ 👍 FP falso positivo	✅ 👎 TN verdadeiro negativo

# Projeto – Naive Bayes

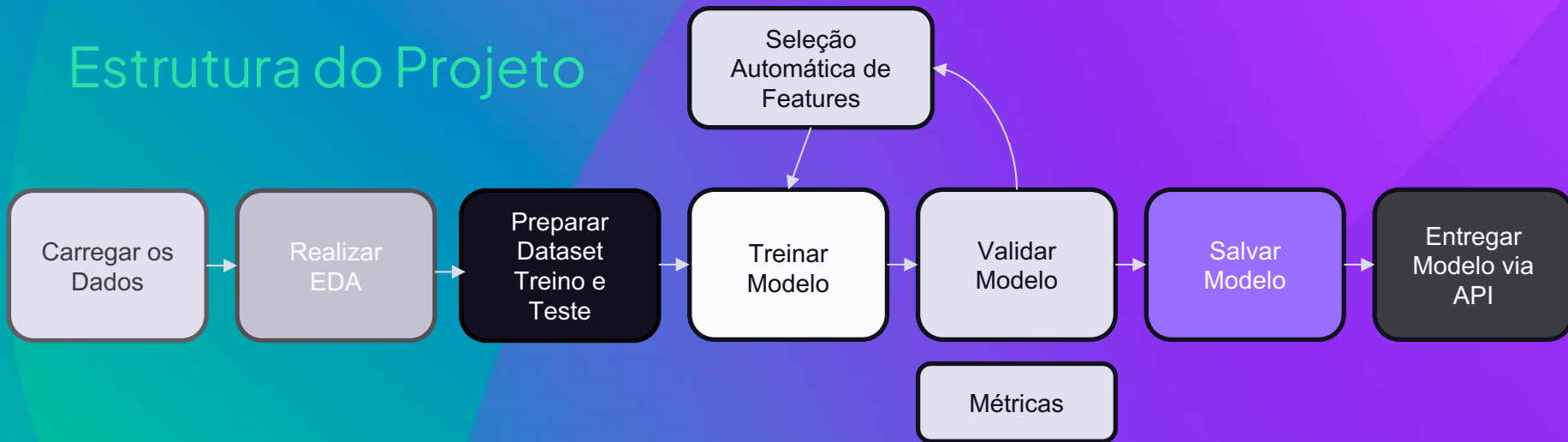
Uma clínica de endocrinologia recebeu dados de um estudo sobre obesidade, onde os respondentes preencheram um formulário com hábitos de saúde (alimentação e exercícios), condições físicas e informações complementares como idade, gênero e histórico familiar. Com base nestas informações, o respondente também preenche se é ou não obeso.

Esta clínica deseja validar se este conjunto de dados poderia ser utilizado para prever se um paciente da clínica é obeso ou não, com base nas respostas que este paciente irá fornecer num formulário similar ao do estudo.

Sendo assim, iremos construir um classificador binário através do algoritmo de Naive Bayes, para definir (prever) se um paciente é ou não obeso, com base nos dados do formulário.

Um aspecto importante é que este questionário não possui informações como peso e altura, o que poderiam já caracterizar de forma mais simples (através do cálculo do IMC), se o paciente pode ser considerado obeso ou não.

# Estrutura do Projeto



# Code Time ...



Rocketseat © 2023  
Todos os direitos reservados

[rocketseat.com.br](https://rocketseat.com.br)

