

# KNN – K Nearest Neighbors

---

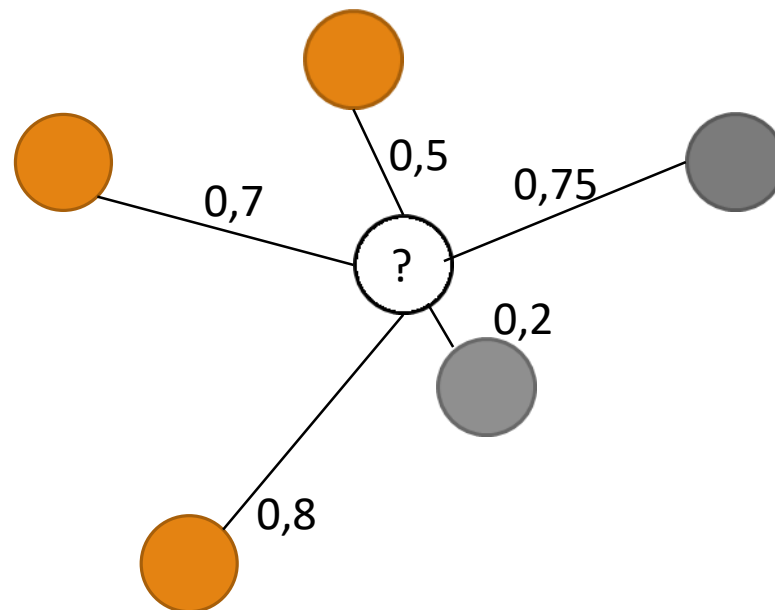
Leonardo De Holanda Bonifácio

[github.com/leothi](https://github.com/leothi)

[lhbonifacio@latam.stefanini.com](mailto:lhbonifacio@latam.stefanini.com)

# Introdução

---

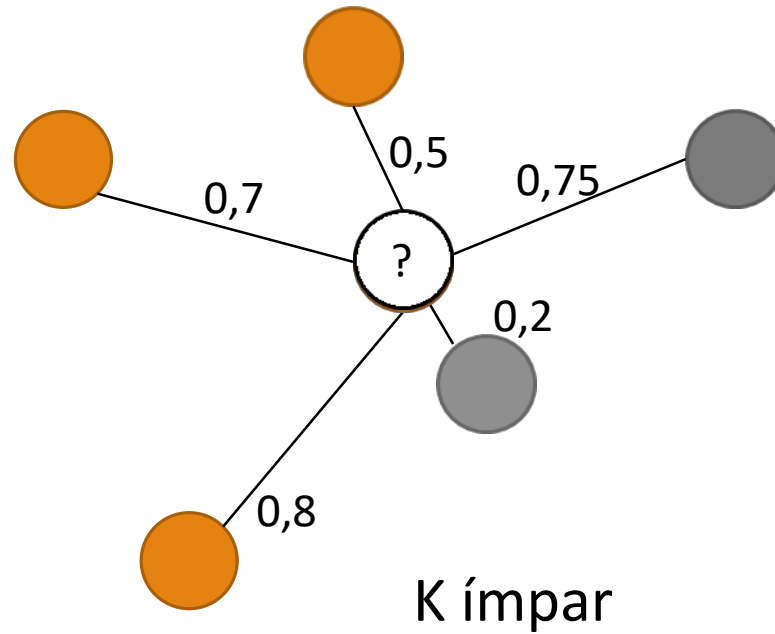


Como classificar um novo registro?

- Distância para todos os vizinhos
- Selecionar K
- Classificar


# Introdução


---



Como classificar um novo registro?

- Distância para todos os vizinhos
- Selecionar K
- Classificar

K = 1 

K = 3 

K = 4  

# K vizinhos mais próximos

---

- Algoritmo de classificação – supervisionado
- Classifica de acordo com os K vizinhos mais próximos ao registro
- Algoritmo Lazy: não salva modelo – previsão ocorre quando uma nova instância precisa ser classificada

# Cálculo da distância

---

- Distância euclidiana:

$$D(a, b) = \sqrt{\sum_i (a_i - b_i)^2}$$

- Exemplo:

	$5 - 5 = 0$	$0^2 = 0$
$a = (5, 7, 9)$	$7 - 5 = 2$	$2^2 = 4$
$b = (5, 5, 5)$	$9 - 5 = 4$	$4^2 = 16$

$$0 + 4 + 16 = 20$$

$$D(a, b) = \sqrt{20} = 4,47$$

$$D(a, b) = \sqrt{\sum_i (a_i - b_i)^2}$$

# Exemplo Completo – Classificação filme

Filme	Violência	Romance	Ação	Comédia	Classe
Anabelle	0,6	0	0,3	0	Terror
A volta dos que não foram	0,9	0	0,5	0,1	Terror
De pernas pro ar 3	0,1	0,2	0,1	0,9	Comédia
Gente grande	0	0,2	0,2	0,8	Comédia

<b>Novo Registro:</b>	<b>Maligno x Anabelle</b>	<b>Maligno x A volta</b>	<b>Maligno x Pernas</b>	<b>Maligno x Gente Grande</b>
	(0,8; 0,1; 0,5; 0,0)	(0,8; 0,1; 0,5; 0,0)	(0,8; 0,1; 0,5; 0,0)	(0,8; 0,1; 0,5; 0,0)
<b>Maligno</b>	(0,6; 0,0; 0,3; 0,0)	(0,9; 0,0; 0,5; 0,1)	(0,1; 0,2; 0,1; 0,9)	(0,0; 0,2; 0,2; 0,8)
Violência = 0,8	$0,2^2 + 0,1^2 + 0,2^2 + 0$	$0,1^2 + 0,1^2 + 0^2 + 0,1^2$	$0,7^2 + 0,1^2 + 0,4^2 + 0,9^2$	$0,8^2 + 0,1^2 + 0,4^2 + 0,8^2$
Romance = 0,1	$0,04 + 0,01 + 0,04 = 0,09$	$0,01 + 0,01 + 0,01 = 0,03$	$0,49 + 0,01 + 0,16 + 0,8 = 1,46$	$0,64 + 0,01 + 0,16 + 0,64 = 1,45$
Ação = 0,5	$D = \sqrt{0,09} = 0,3$	$D = \sqrt{0,03} = 0,17$	$D = \sqrt{1,46} = 1,2$	$D = \sqrt{1,45} = 1,2$
Comédia = 0				

# Resultado

---

- Ordem das distâncias:

1.	0,17 - A volta: <b>Terror</b>	<b>K = 1</b>	1 Terror: <b>Terror</b>
2.	0,3 - Anabelle: <b>Terror</b>	<b>K = 2</b>	2 Terror: <b>Terror</b>
3.	1,2 - Pernas: <b>Comédia</b>	<b>K = 3</b>	2 Terror, 1 Comédia: <b>Terror</b>
4.	1,2 - Gente: <b>Comédia</b>	<b>K = 4</b>	2 Terror, 2 Comédia: <b>Terror/Comédia</b>

# Observações

---

- Normalização:

Idade	Renda	Classe
30	50000	Pobre
60	150000	Rico

- Simples, poderoso e rápido
- Relacionamento entre previsores é complexo
- K pequeno – prejudicado com ruídos e outliers
- K grande – overfitting
- K = 3 ou 5



# Implementação

---

- Biblioteca em Python: scikit-learn
- Carregamento da base -> Pré-processamento -> Classe e método fit -> Resultados
- Código de exemplo