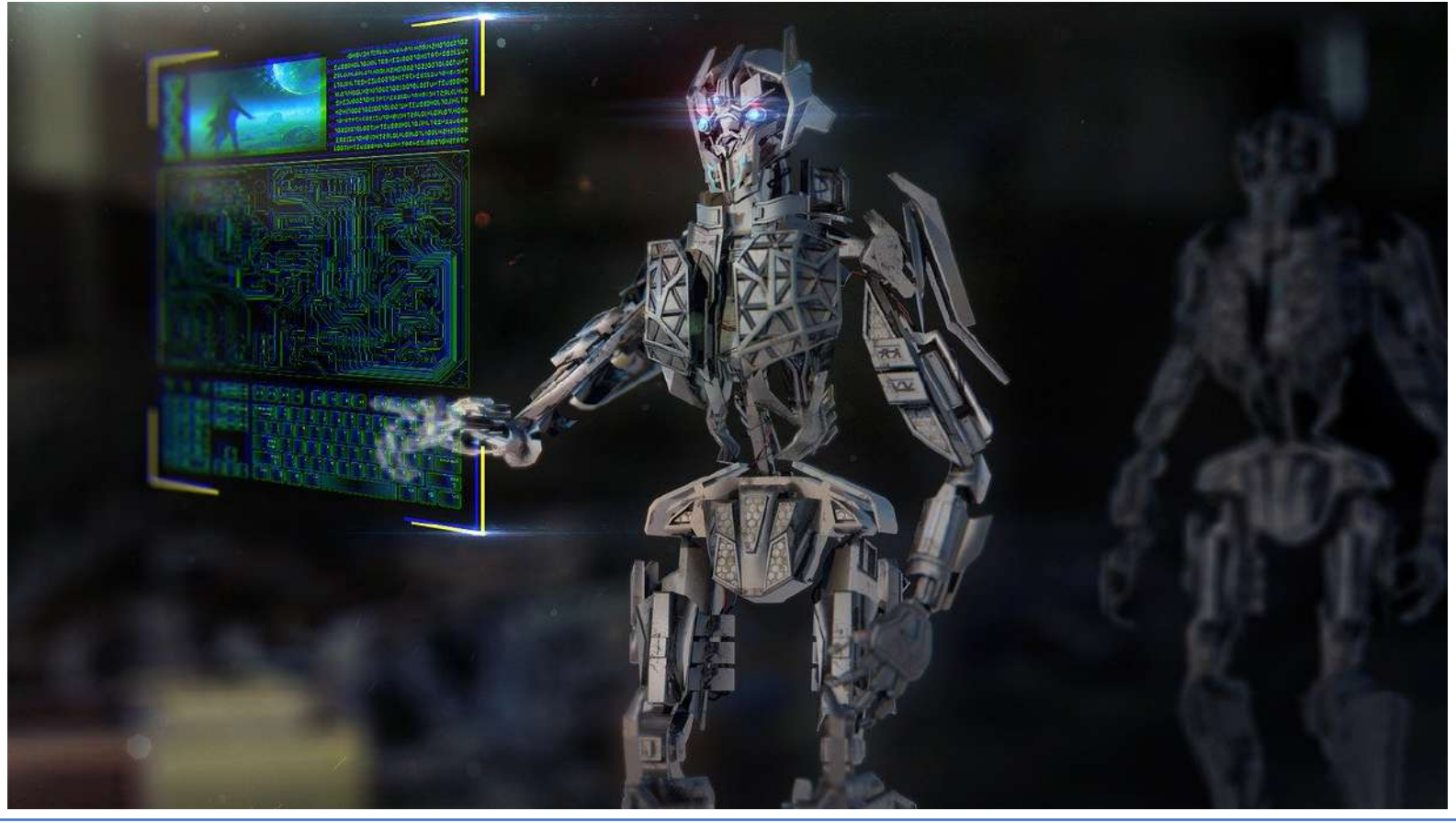
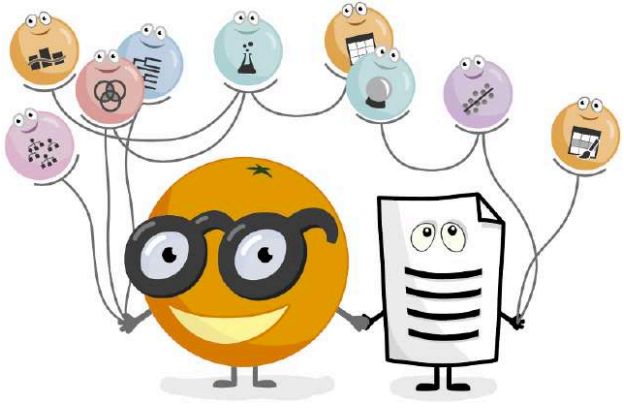
MACHINE LEARNING E DATA SCIENCE: O GUIA PARA INICIANTES



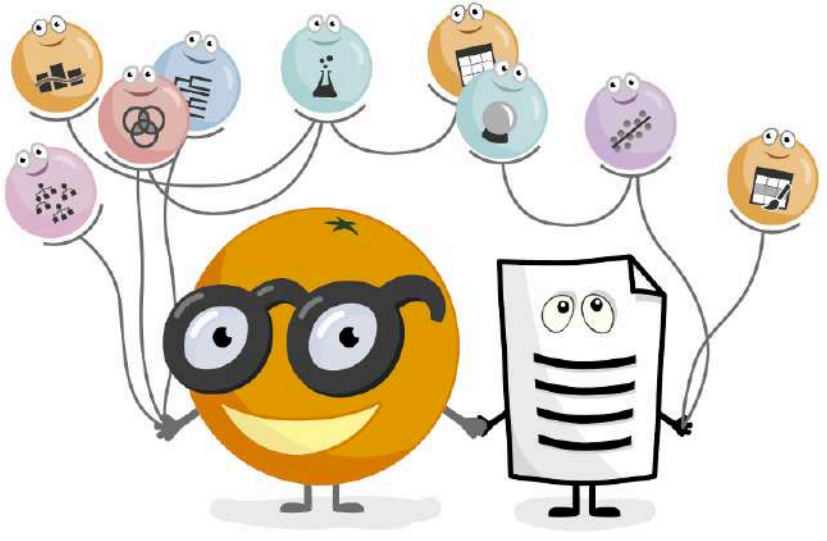
CONTEÚDO



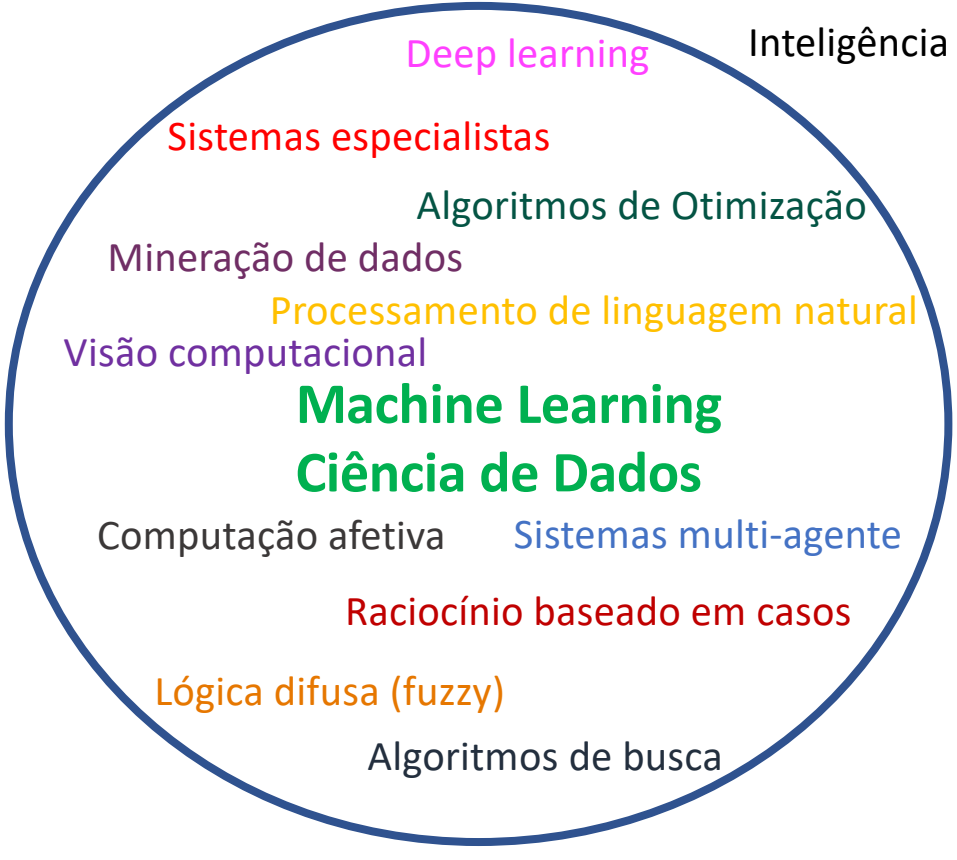
- Conceitos sobre Machine Learning
- Classificação
  - Naïve Bayes
  - Árvores de decisão
  - Aprendizagem por regras
  - Aprendizagem baseada em instâncias (kNN)
  - Aprendizagem de máquinas de vetores de suporte (SVM)
  - Redes Neurais Artificiais
- Avaliação de algoritmos
- Implementações com Orange

CONTEÚDO

- Regressão
  - Regressão linear
- Agrupamento
  - Algoritmo k-means
- Associação
  - Algoritmo apriori
- Implementações com Orange



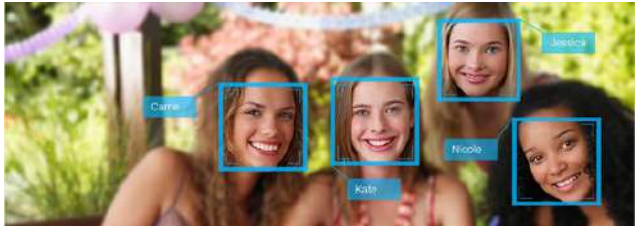
MACHINE LEARNING E DATA SCIENCE



Machine Learning

Classificação
Regressão
Agrupamento
Associação
Aprendizagem por reforço

MACHINE LEARNING E DATA SCIENCE



Google

facebook Ads

eHarmony®

NETFLIX

amazon.com

EXPERT ACADEMY

CLASSIFICAÇÃO

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	< 15.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

Treinamento

História do crédito	Dívida	Garantias	Renda anual
Ruim	Alta	Adequada	< 15.000
Desconhecida	Alta	Adequada	< 15.000
Desconhecida	Baixa	Nenhuma	> 35.000
Boa	Alta	Adequada	>= 15.000 a <= 35.000

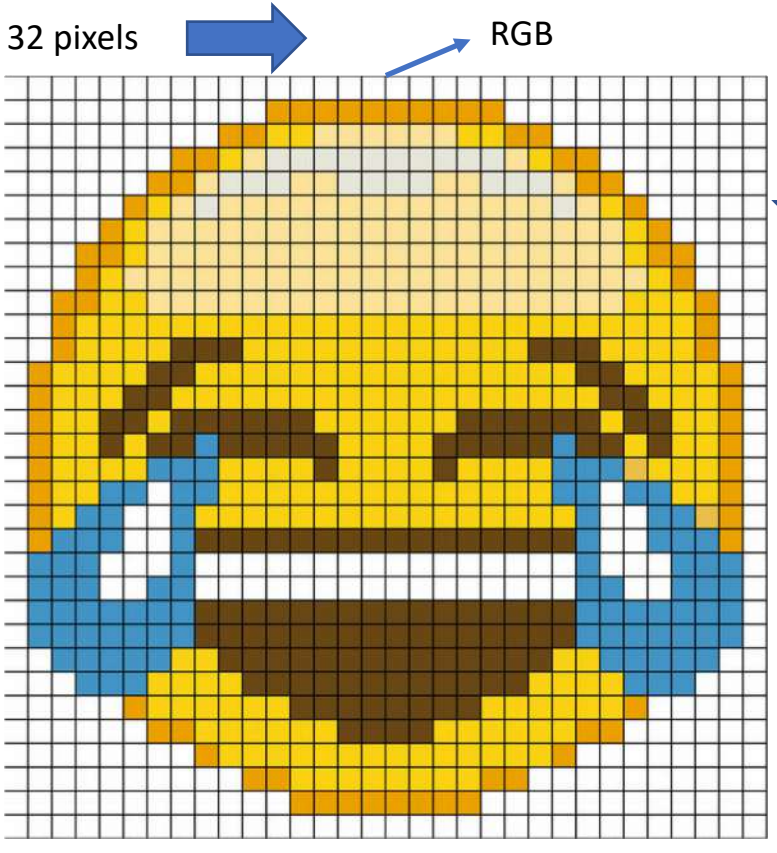
CLASSIFICAÇÃO

Sexo	País	Idade	Comprar
M	França	25	Sim
M	Inglaterra	21	Sim
F	França	23	Sim
F	Inglaterra	34	Sim
F	França	30	Não
M	Alemanha	21	Não
M	Alemanha	20	Não
F	Alemanha	18	Não
F	França	34	Não
F	França	34	Não
M	França	55	Não
M	Inglaterra	25	Sim
M	Alemanha	48	Sim
F	Inglaterra	23	Não

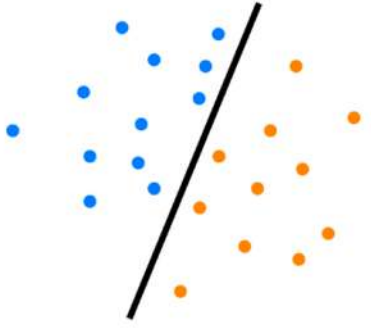
Treinamento

Sexo	País	Idade
M	França	38
F	Inglaterra	25
M	Alemanha	55
F	França	20

CLASSIFICAÇÃO



32 x 32 = 1024 **pixels**



NAÏVE BAYES

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	< 15.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

História = Boa
Dívida = Alta
Garantias = Nenhuma
Renda = > 35

Soma: 0,0079 + 0,0052 + 0,0514 = **0,0645**

P(Alto) = 6/14 * 1/6 * 4/6 * 6/6 * 1/6
P(Alto) = 0,0079
P(Alto) = 0,0079 / 0,0645 * 100 = **12,24%**

P(Moderado) = 3/14 * 1/3 * 1/3 * 2/3 * 1/3
P(Moderado) = 0,0052
P(Moderado) = 0,0052 / 0,0645 * 100 = **8,06%**

P(Baixo) = 5/14 * 3/5 * 2/5 * 3/5 * 5/5
P(Baixo) = 0,0514
P(Baixo) = 0,0514 / 0,0645 * 100 = **79,68%**

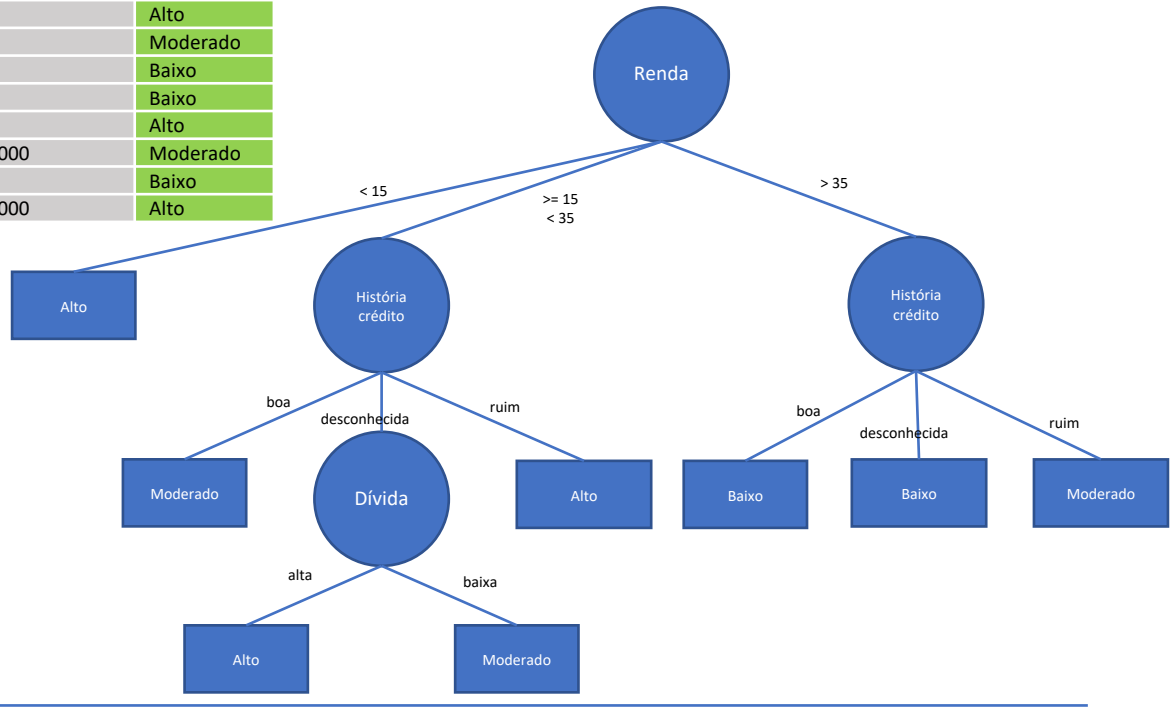
Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 5	Ruim 4	Alta 7	Baixa 7	Nenhuma 11	Adequada 3	< 15000 3	>= 15000 <= 35000 4	> 35000 7
Alto 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
Moderado 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
Baixo 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

ÁRVORE DE DECISÃO

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	< 15.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

História = Boa
Dívida = Alta
Garantias = Nenhuma
Renda = > 35

História = Ruim
Dívida = Alta
Garantias = Adequada
Renda = < 15



APRENDIZAGEM POR REGRAS

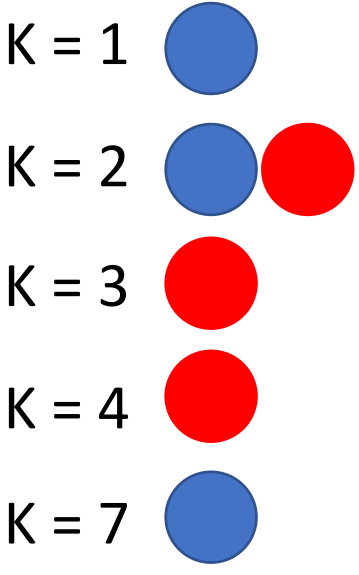
História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	< 15.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

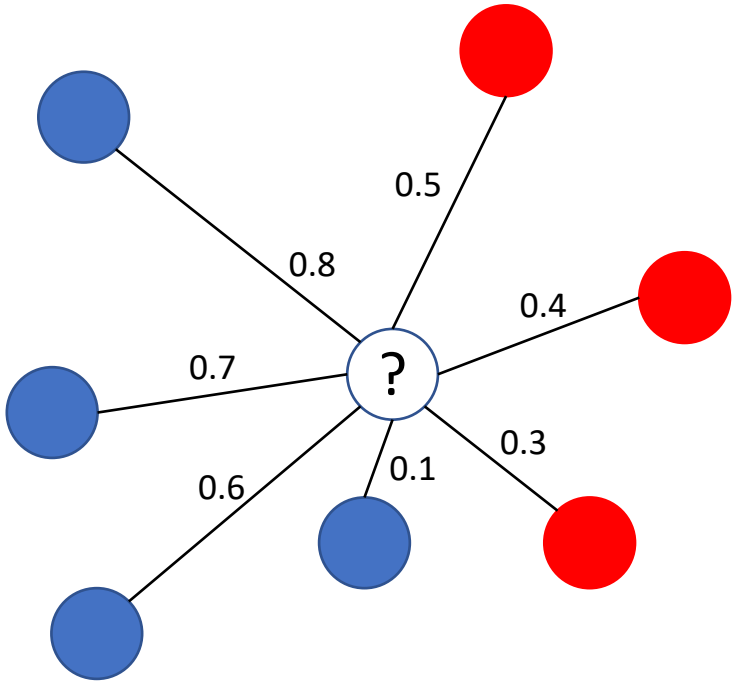
História = Boa
Dívida = Alta
Garantias = Nenhuma
Renda => 35

História = Ruim
Dívida = Alta
Garantias = Adequada
Renda =< 15

Regra	Resultado
Se renda =>35.000 E história de crédito = BOA	Risco = BAIXO
Se renda =>35.000 e história de crédito = DESCONHECIDA	Risco = BAIXO
Default (padrão)	Risco = ALTO

APRENDIZAGEM BASEADA EM INSTÂNCIAS - kNN





$$DE(x,y) = \sqrt{\sum_{i}^{p}(x_i - y_i)^2}$$

- x = 5, 7, 9
- y = 5, 5, 5
- Subtração de cada posição do vetor
  - 5 – 5 = 0
  - 7 – 5 = 2
  - 9 – 5 = 4
- Elevação ao quadrado
  - $0^2 = 0$
  - $2^2 = 4$
  - $4^2 = 16$
- Somatório
  - 0 + 4 + 16 = 20
- Raiz quadrada
  - Raiz(20) = 4,47
- **Distância Euclidiana = 4,47**

APRENDIZAGEM BASEADA EM INSTÂNCIAS - kNN

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	< 15.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

História = Boa (1)
Dívida = Alta (1)
Garantias = Nenhuma (1)
Renda = > 35 (3)

Novo x 9º

1 1 1 3
1 2 1 3

$0 + 1^2 + 0 + 0$
$0 + 1 + 0 + 0 = 1$
Raiz(1) = **1**

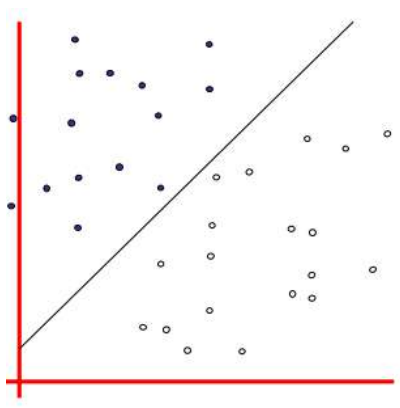
Novo x 3º

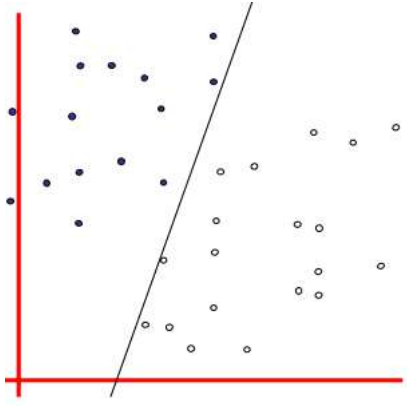
1 1 1 3
2 2 1 2

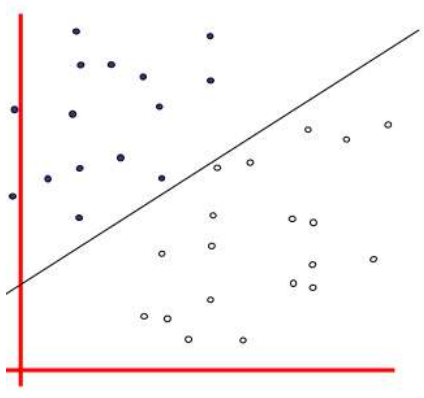
$1^2 + 1^2 + 0 + 1^2$
$1 + 1 + 0 + 1 = 3$
Raiz(3) = **1,7**

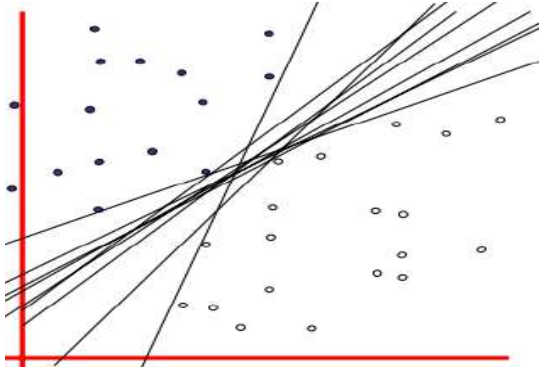
História do crédito	Dívida	Garantias	Renda anual	Risco
3	1	1	1	Alto
2	1	1	2	Alto
2	2	1	2	Moderado
2	2	1	3	Alto
2	2	1	3	Baixo
2	2	2	3	Baixo
3	2	1	1	Alto
3	2	2	3	Moderado
1	2	1	3	Baixo
1	1	2	3	Baixo
1	1	1	1	Alto
1	1	1	2	Moderado
1	1	1	3	Baixo
3	1	1	2	Alto

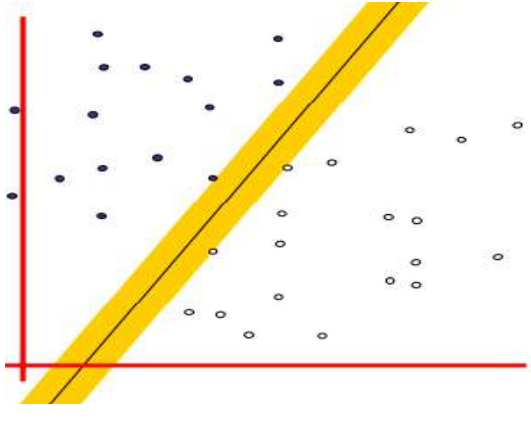
APRENDIZAGEM COM MÁQUINAS DE VETORES DE SUPORTE (SVM)

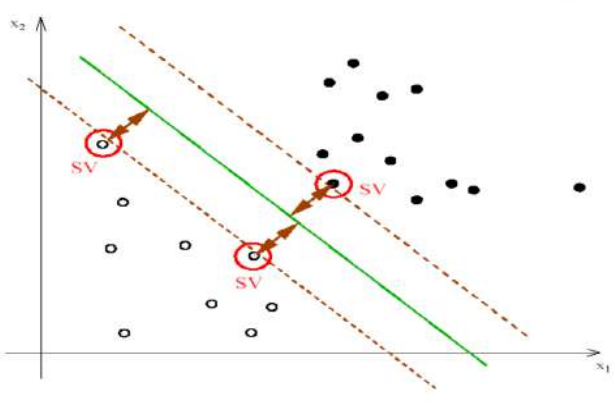




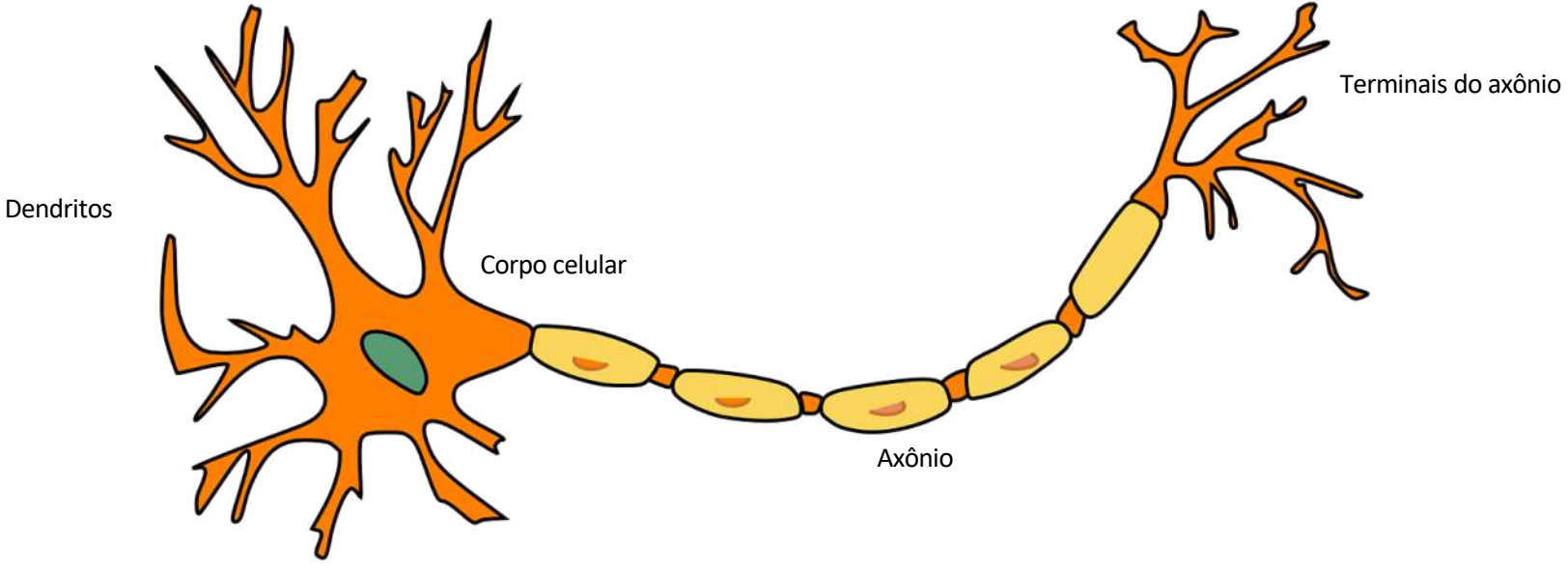




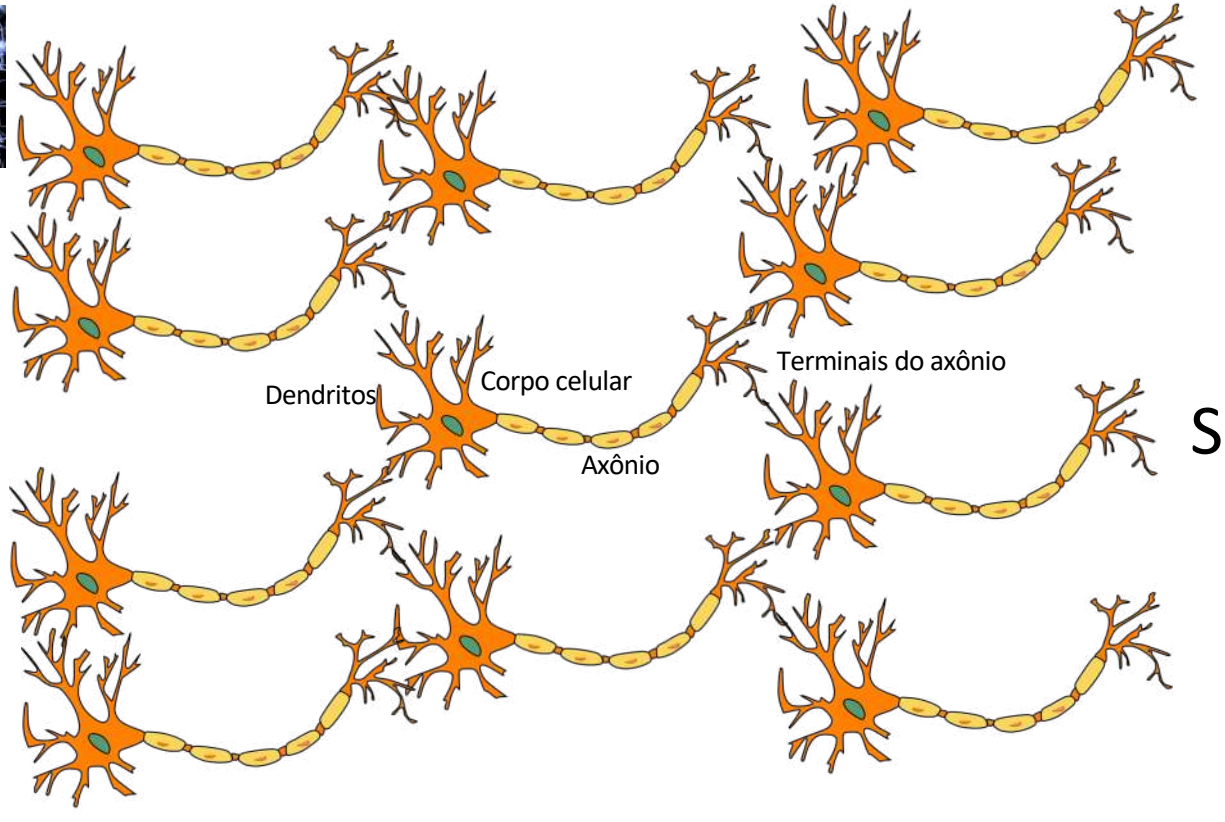




REDES NEURAIS ARTIFICIAIS

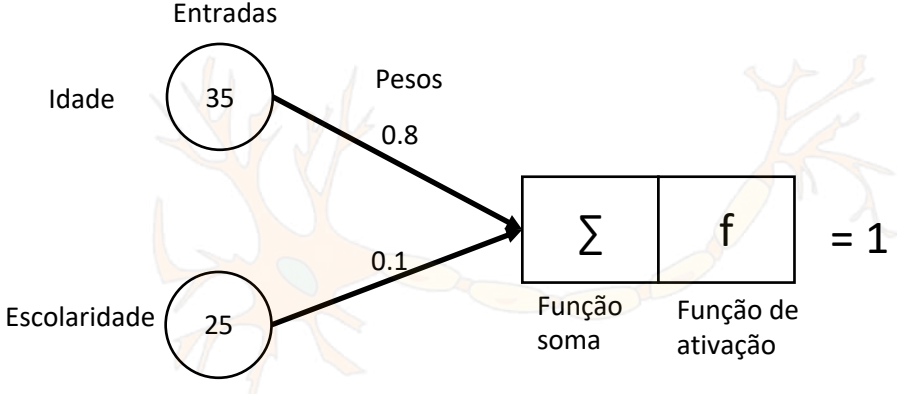


REDES NEURAIS ARTIFICIAIS



Sinapse

REDES NEURAIS ARTIFICIAIS – PERCEPTRON



$$soma = \sum_{i=1}^{n} xi * wi$$

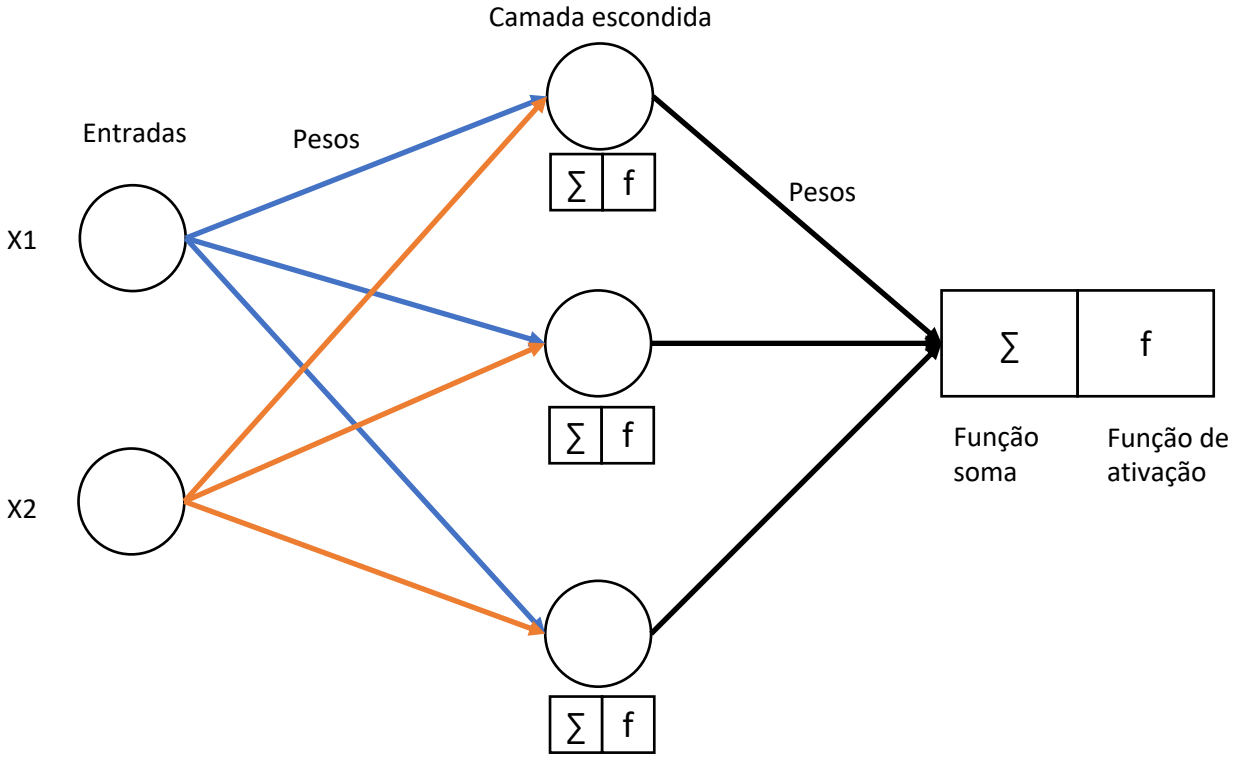
$$soma = (35 * 0.8) + (25 * 0.1)$$

$$soma = 28 + 2.5$$

$$soma = 30.5$$

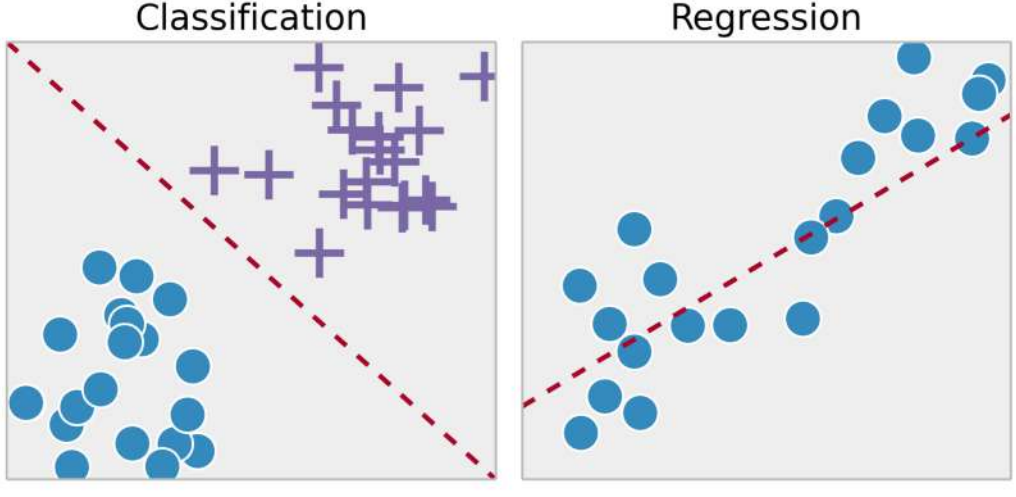
Maior ou igual a 1 = 1
Caso contrário = 0

REDES NEURAIS ARTIFICIAIS – REDES MULTICAMADA

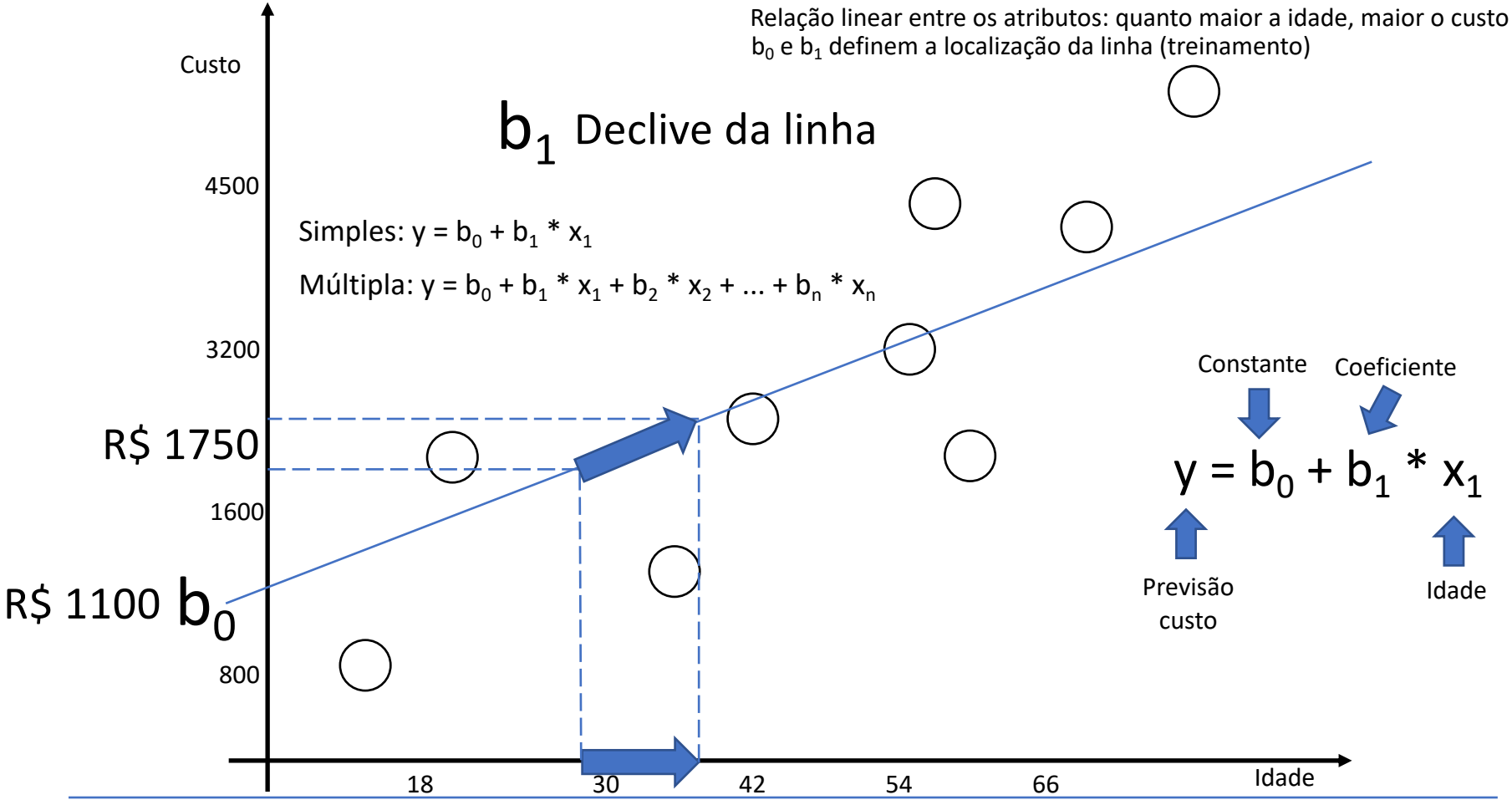


REGRESSÃO

- Gastos propaganda (x) -> valor de venda (y)
- Temperatura, umidade e pressão do ar (x) -> velocidade do vento (y)
- Fatores externos (x) -> valor do dólar (y)
- Resultados do exame (x) -> probabilidade de um paciente sobreviver (y)
- Risco de investimento
- Gastos no cartão de crédito, histórico (x) -> limite (y)
- Valores anteriores (x) -> valores de produtos (y)

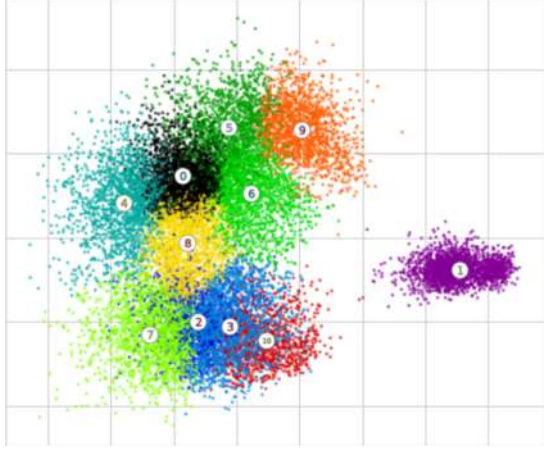


REGRESSÃO LINEAR

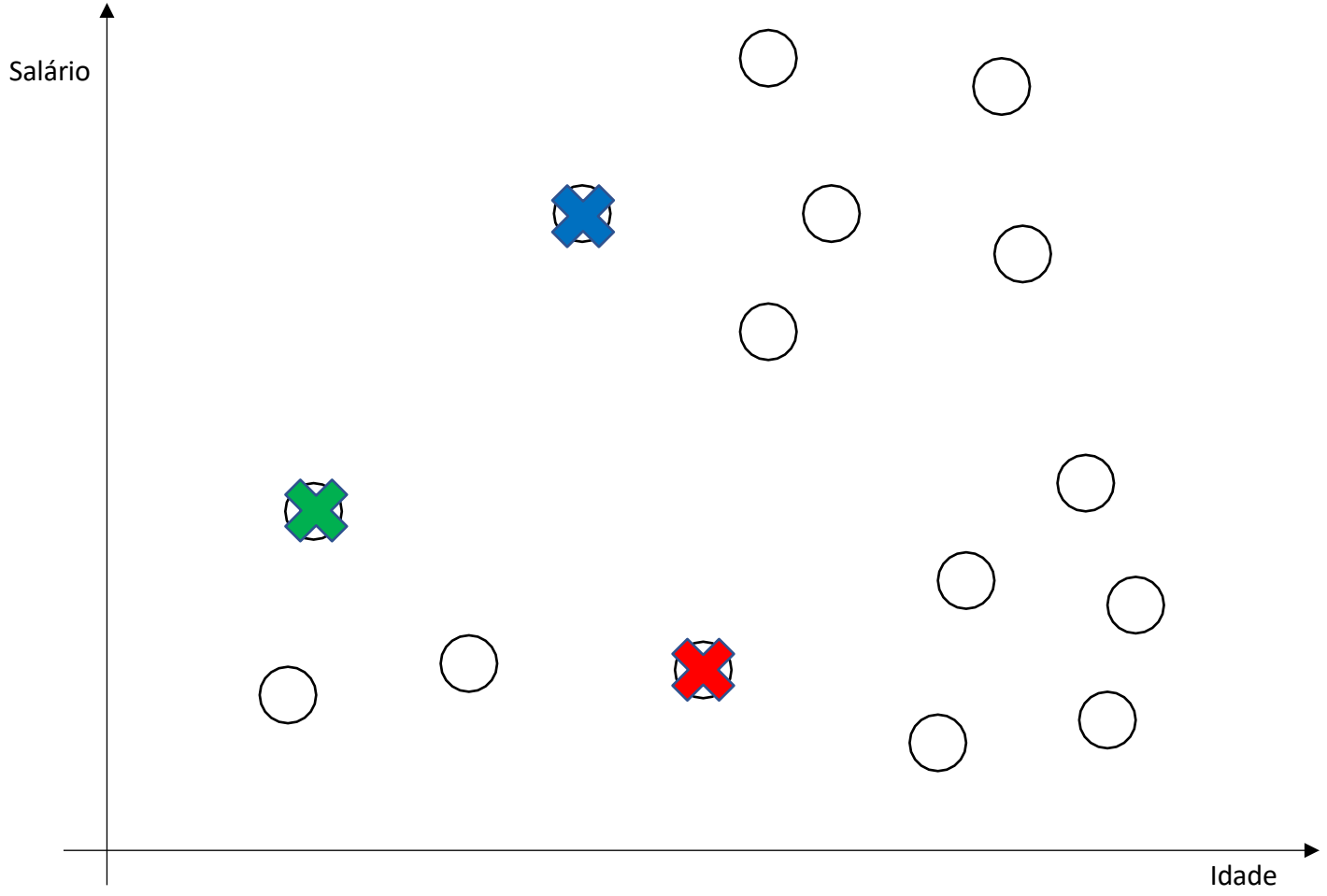


AGRUPAMENTO

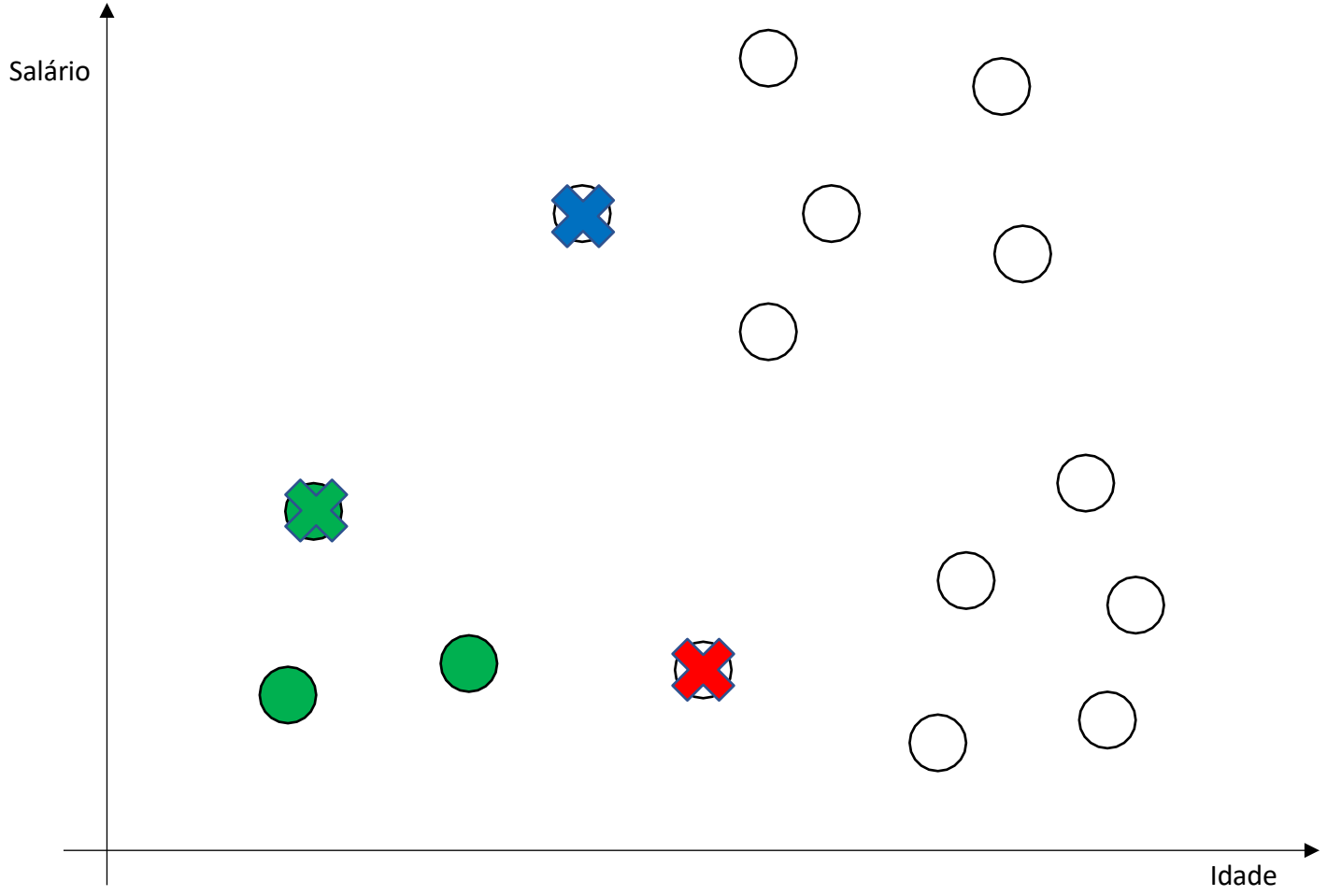
- Segmentação de mercado
- Encontrar grupos de clientes que irão comprar um produto (mala direta)
- Agrupamento de documentos/notícias
- Agrupamento de produtos similares
- Perfis de clientes (NetFlix)
- Análise de redes sociais



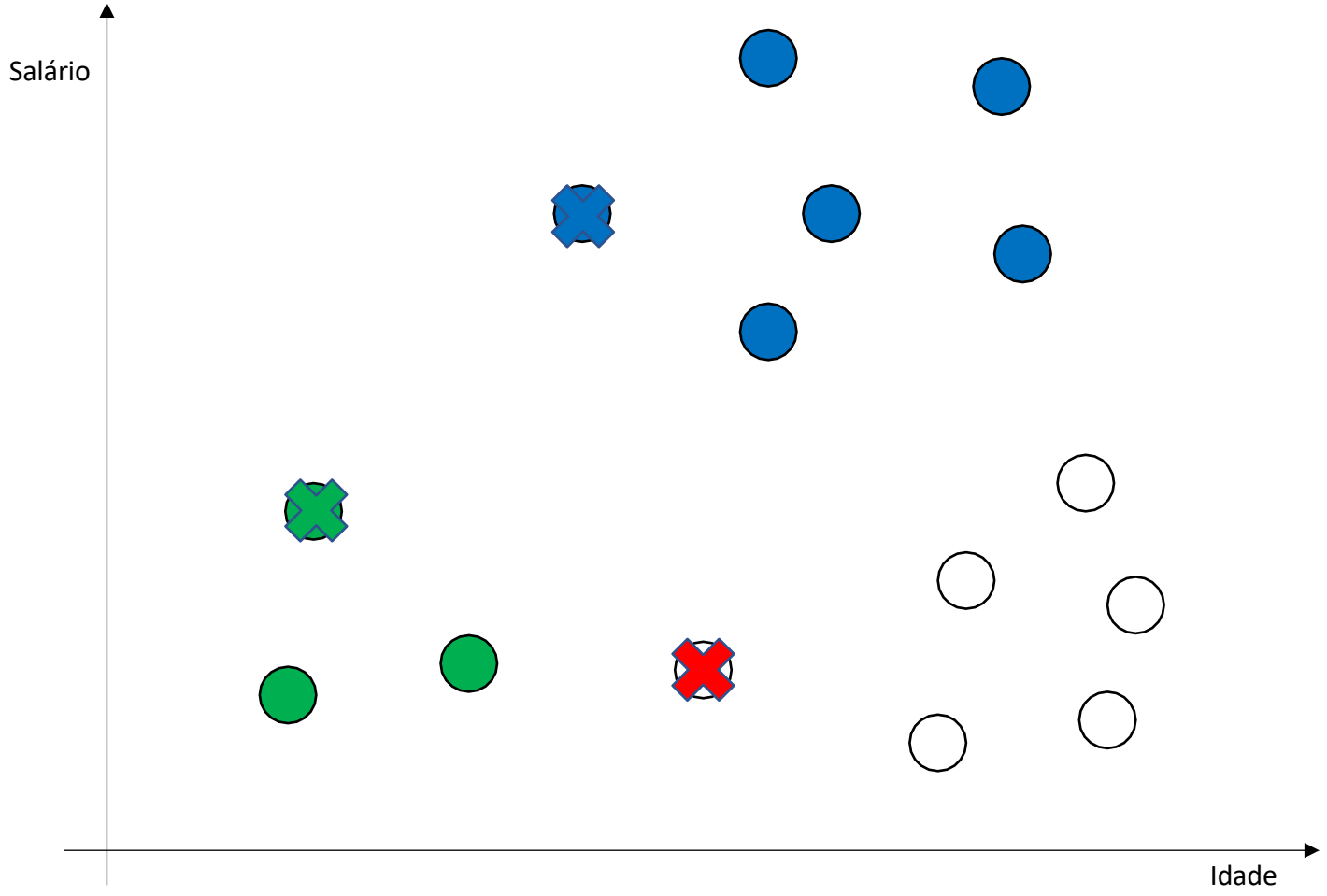
AGRUPAMENTO COM K-MEANS



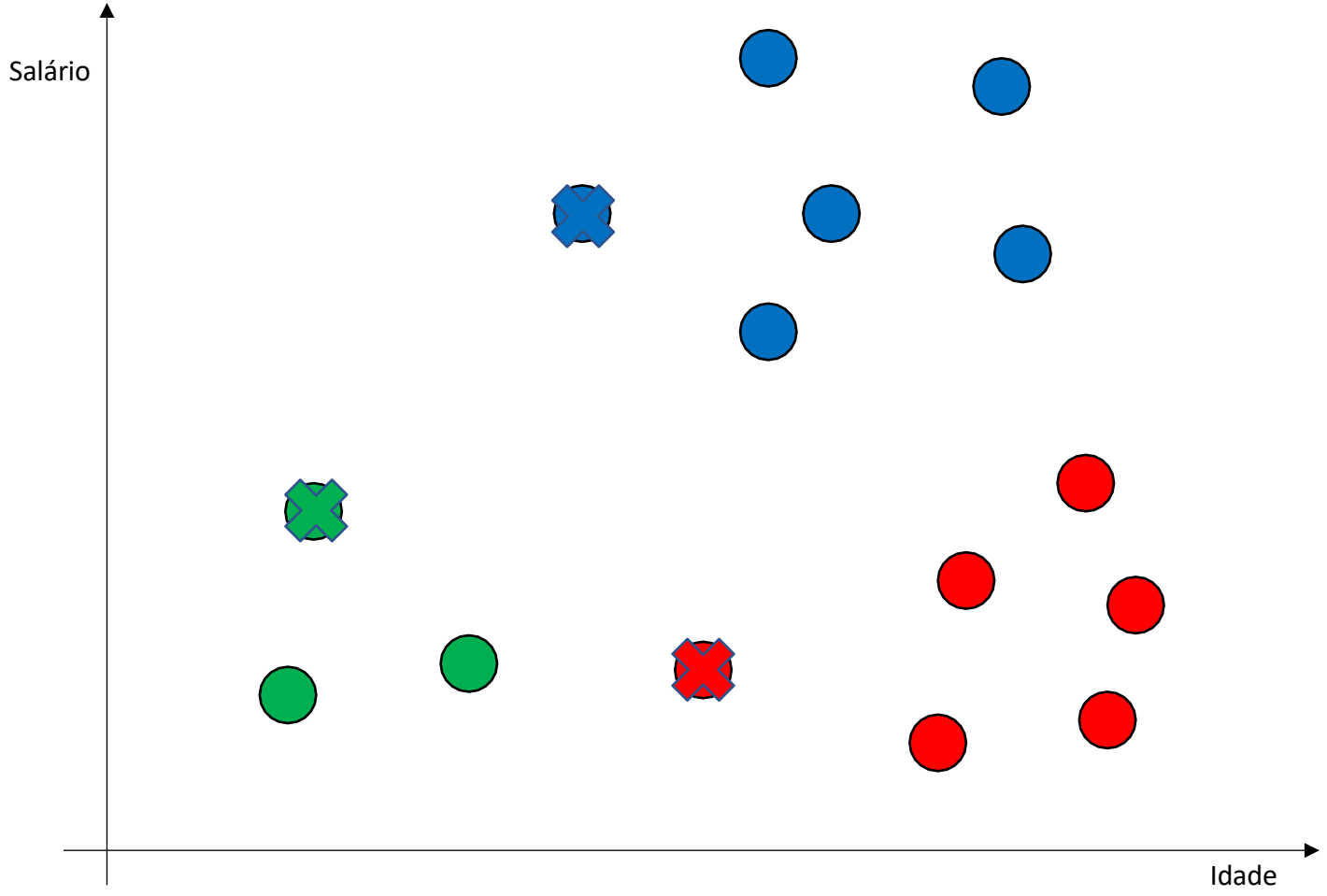
AGRUPAMENTO COM K-MEANS



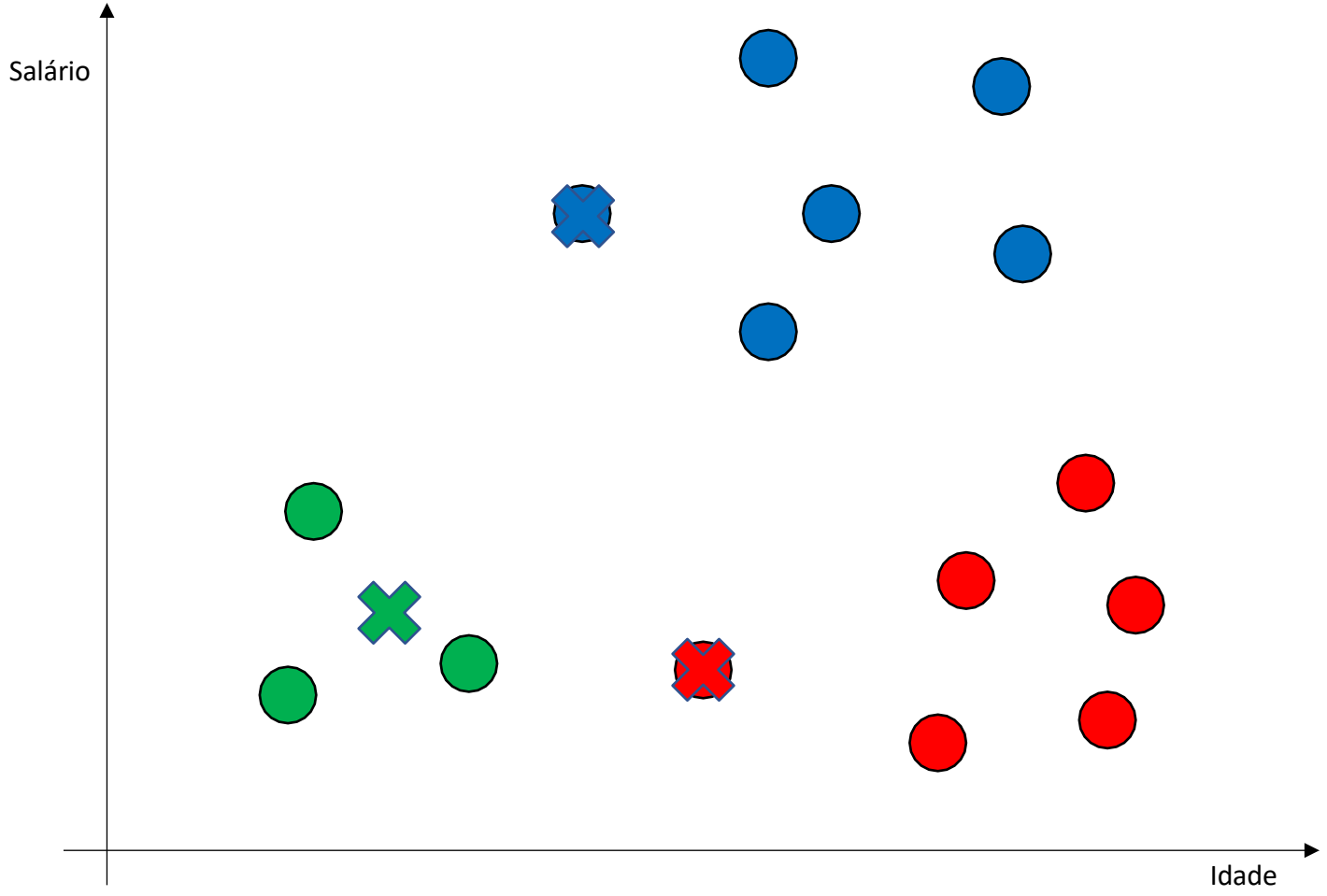
AGRUPAMENTO COM K-MEANS



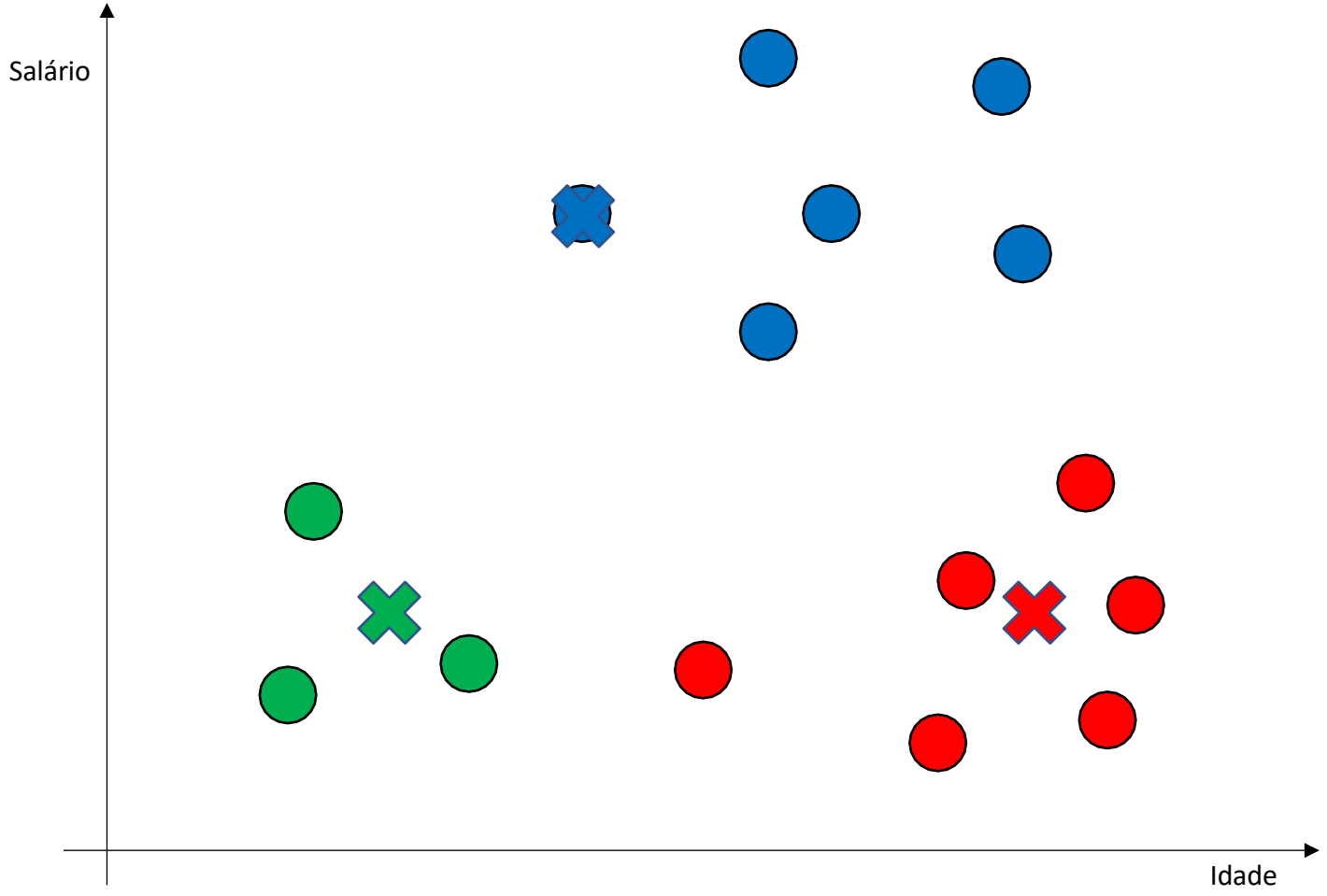
AGRUPAMENTO COM K-MEANS



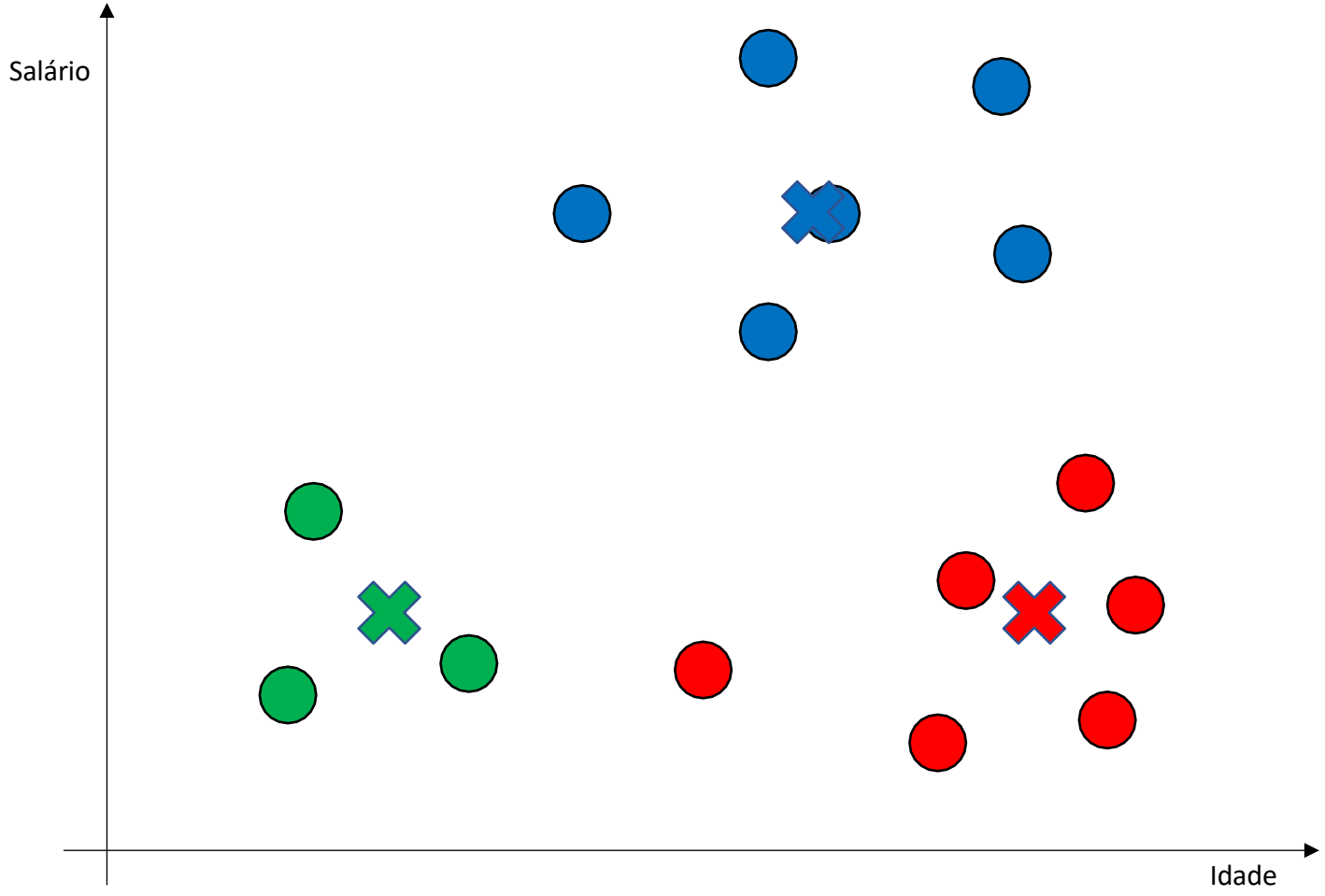
AGRUPAMENTO COM K-MEANS



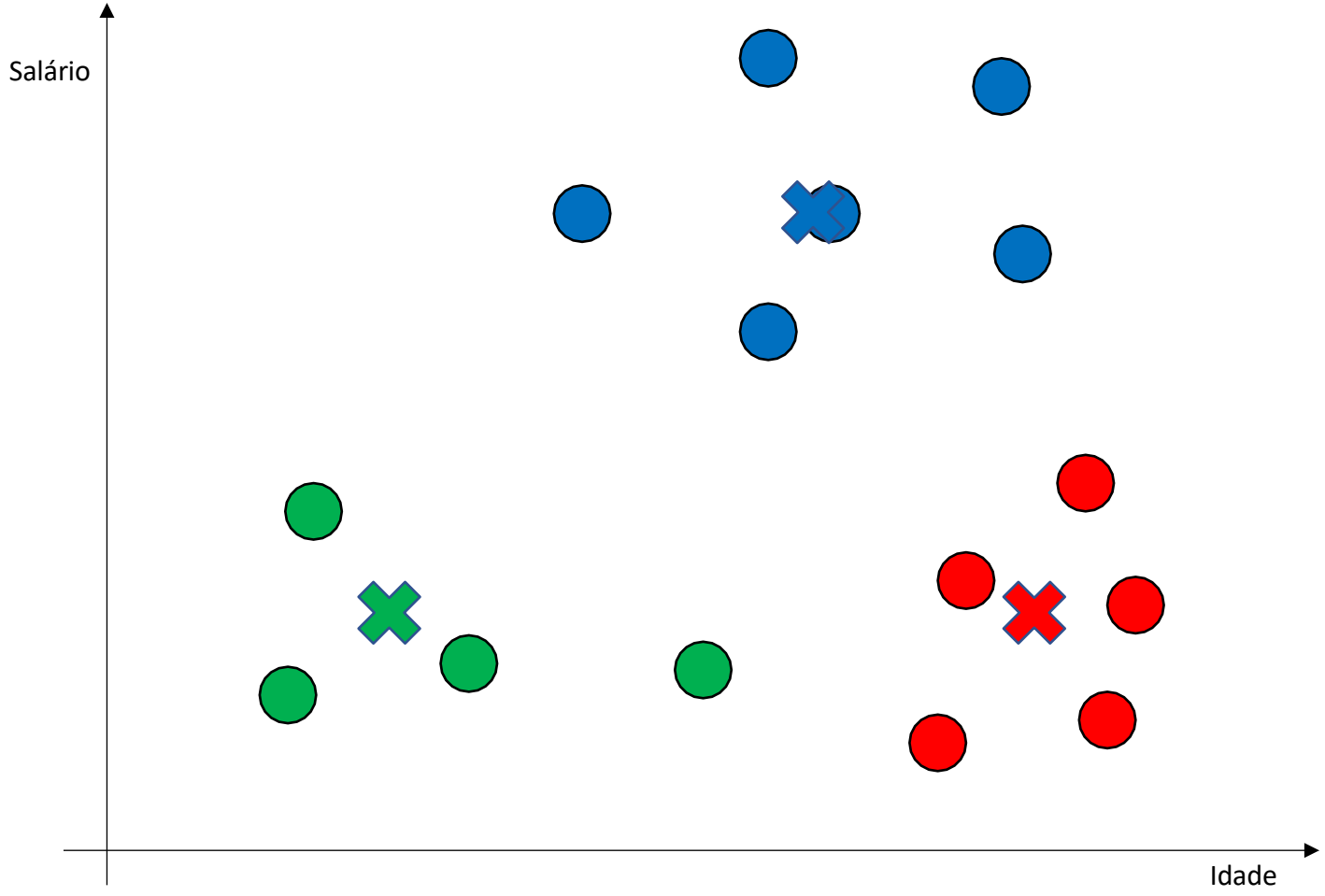
AGRUPAMENTO COM K-MEANS



AGRUPAMENTO COM K-MEANS

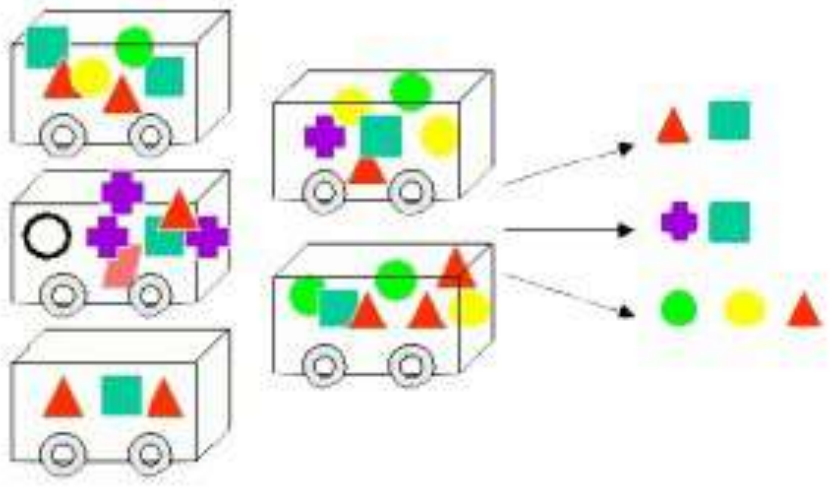


AGRUPAMENTO COM K-MEANS



ASSOCIAÇÃO

- Prateleiras de mercado
  - Em que prateleira o biscoito de chocolate deve ser colocado para maximizar suas vendas?
  - Suco de uva costuma ser comprado com refrigerante?
  - Qual produto pode ser colocado em promoção para uma venda casada com tomates?
- Promoções com itens que são vendidos em conjunto
- Planejar catálogos das lojas e folhetos de promoções
- Controle de evasão em universidades



ALGORITMO APRIORI

Nº	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	Não	Sim	Não	Sim	Sim	Não	Não
2	Sim	Não	Sim	Sim	Sim	Não	Não
3	Não	Sim	Não	Sim	Sim	Não	Não
4	Sim	Sim	Não	Sim	Sim	Não	Não
5	Não	Não	Sim	Não	Não	Não	Não
6	Não	Não	Não	Não	Sim	Não	Não
7	Não	Não	Não	Sim	Não	Não	Não
8	Não	Não	Não	Não	Não	Não	Sim
9	Não	Não	Não	Não	Não	Sim	Sim
10	Não	Não	Não	Não	Não	Sim	Não

ALGORITMO APRIORI

Nº	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	Não	Sim	Não	Sim	Sim	Não	Não
2	Sim	Não	Sim	Sim	Sim	Não	Não
3	Não	Sim	Não	Sim	Sim	Não	Não
4	Sim	Sim	Não	Sim	Sim	Não	Não
5	Não	Não	Sim	Não	Não	Não	Não
6	Não	Não	Não	Não	Sim	Não	Não
7	Não	Não	Não	Sim	Não	Não	Não
8	Não	Não	Não	Não	Não	Não	Sim
9	Não	Não	Não	Não	Não	Sim	Sim
10	Não	Não	Não	Não	Não	Sim	Não

Confiança = Número de registros com X e Y / Número total de registros com X

Confiança >= 0,8

{café, pão}

SE café ENTÃO pão – confiança = 3 / 3 = 1,0
SE pão ENTÃO café – confiança = 3 / 5 = 0,6

Suporte = Número de registros com X e Y / Número total de registros

Suporte >= 0,3
Café
Pão
Manteiga