

# ART|IEEEFICIAIS



A stylized illustration of a human brain, split vertically. The left hemisphere is purple and the right is blue. Grey circuit lines with circular nodes extend from the brain, resembling a neural network or computer circuitry.

# Machine Learning

**Regressão, Classificação e Clustering**



**Carine Gottschall**

**Lucas Alves**



# Conteúdo Programático

1. Gestão de pacotes e ambientes em Python
  1. Anaconda
  2. Jupyter Notebook
  3. Google Colab
2. Pacotes essenciais ao desenvolvimento de RNA com Python
  1. Numpy
  2. Pandas
  3. Tratamento de Dados
3. Machine Learning
  1. Regressão Linear
  2. Classificação
  3. Clustering (K-means)

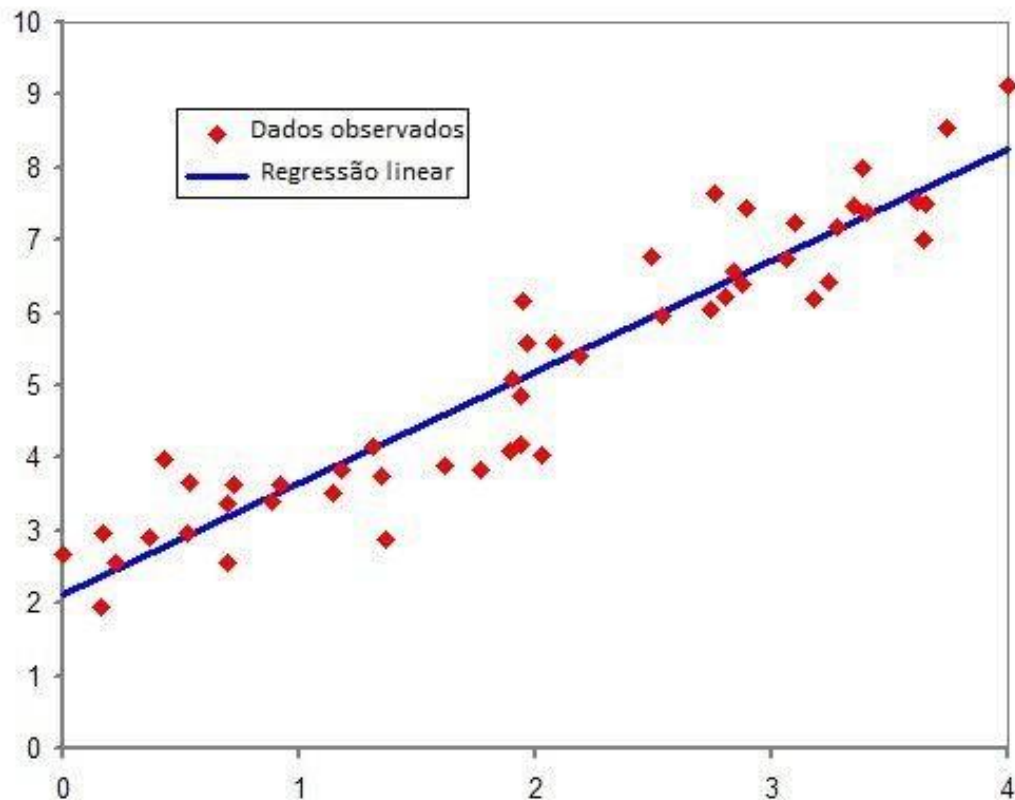
ARTIFICIAIS

# O que é regressão linear?

## Correlação linear

- A verificação da existência de um relacionamento entre duas variáveis. **Dado X e Y, quanto que X explica Y.** Para isso, a regressão linear utiliza os pontos de dados para encontrar a melhor linha de ajuste para modelar essa relação.

- ❑ O resultado da regressão linear é sempre um número. É utilizada adequadamente quando o dataset apresenta algum tipo de tendência de crescimento/decrescimento constante.



A linha traçada pode ser representada pela equação,  $y_i = \alpha + \beta X_i + \varepsilon_i$ , onde  $y$  é a variável explicada (dependente) e representa o que o modelo tentará prever;  $\alpha$  é a constante, representa a interceptação da reta com o eixo vertical;  $\beta$  representa a inclinação em relação à variável explicativa;  $X$  é a variável explicativa (independente) e  $\varepsilon$  representa os valores residuais e possíveis erros.

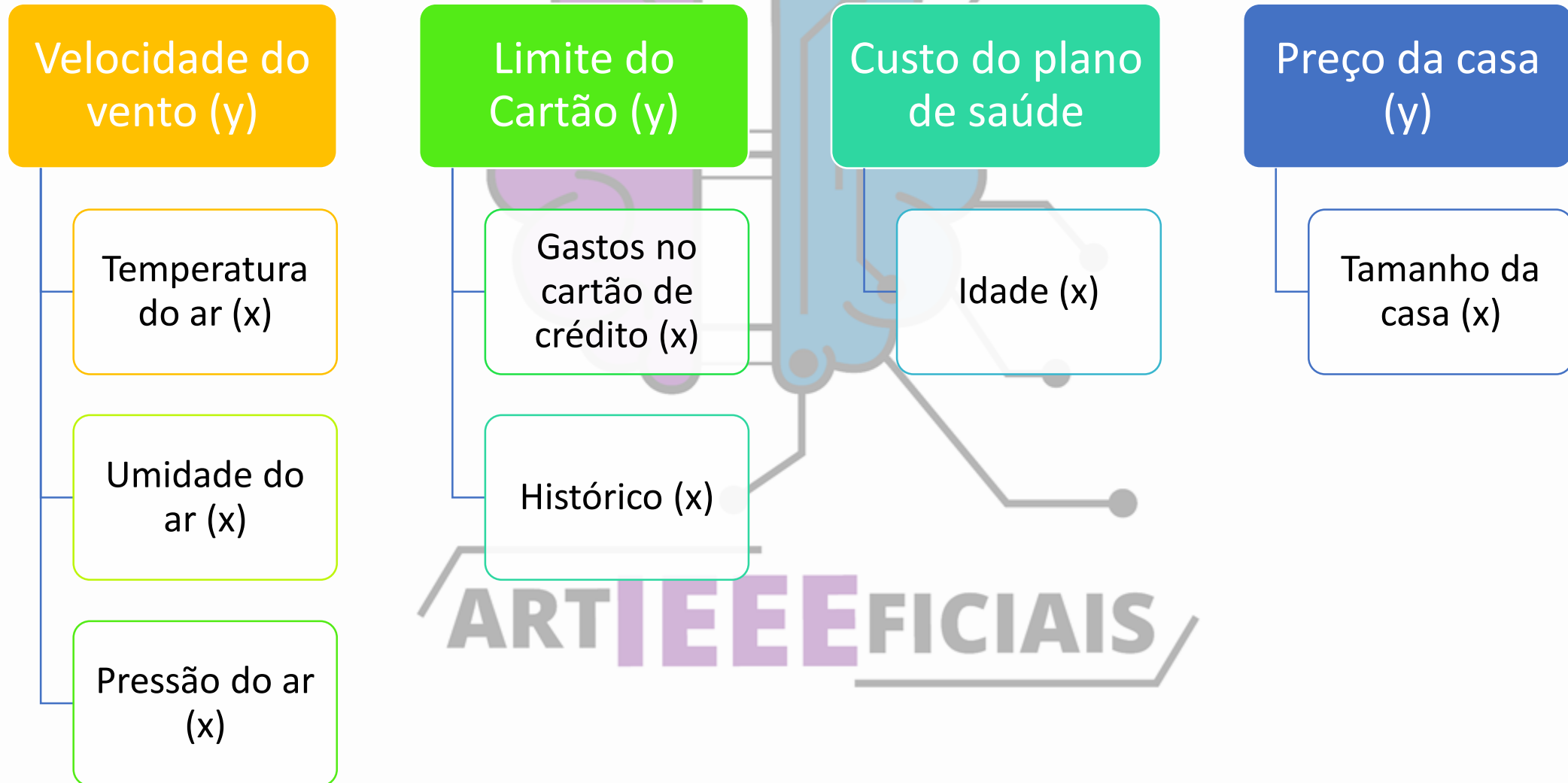
# Coeficiente de correlação



- ❖ Mede o grau da correlação e a direção dessa correlação (positiva ou negativa).
- ❖ Quanto mais próximo dos extremos maior o grau de correlação e quanto mais próximo a zero menor o grau de correlação.

Devemos levar em conta o *grau de correlação* em nossas previsões. Geralmente **acima de 70%** consideramos uma correlação significativa.

# Previsões com regressão



A stylized illustration of a brain with circuit-like lines extending from it, positioned behind a blue ribbon banner.

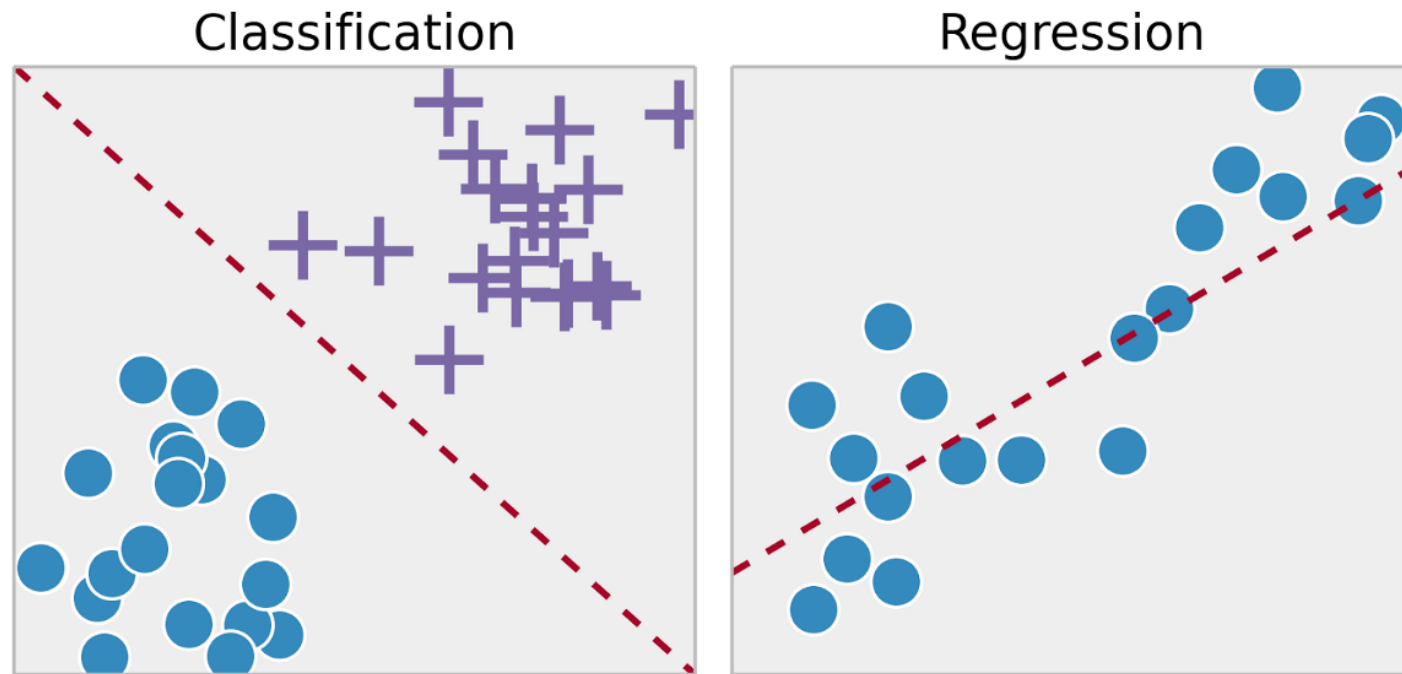
**Hora da prática**

**ART|EEEEFICIAIS**



# O que é Classificação?

- ❑ Classificação é o processo de tomar algum tipo ou conjunto de atributos de entrada e atribuir um rótulo a ela.



# Onde aplicar?

- ❑ Sistemas de classificação são usados geralmente quando as previsões são de natureza distinta, ou seja, um simples “sim ou não”.

*Exemplo:* Mapeamento de uma imagem de uma pessoa e classificação como masculino ou feminino.

ART|EEEEFI

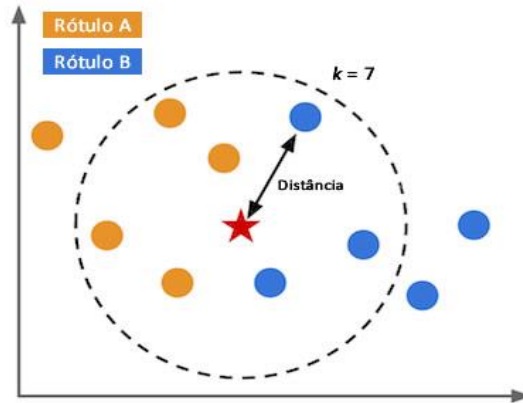
# Classificação supervisionada

❖ O treinamento é feito através de um conjunto de dados pré-definido.



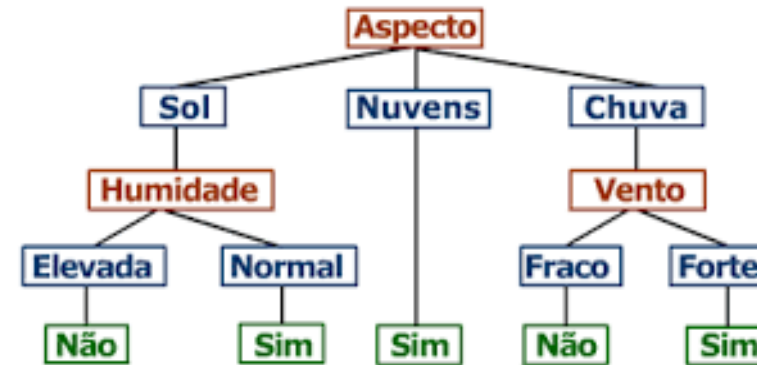
# Métodos de Classificação

## KNN



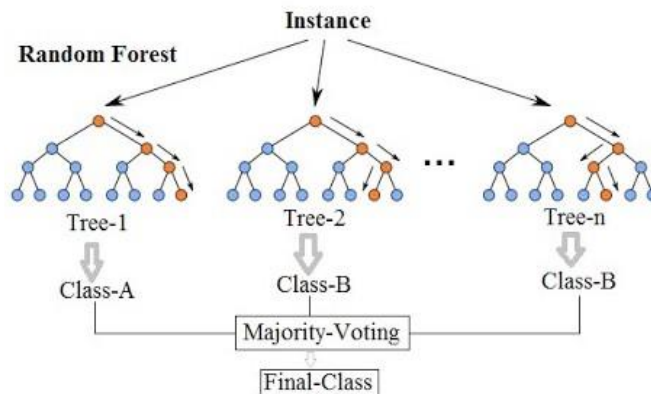
## Árvore de Decisão

### Árvore de Decisão para Jogar Tênis

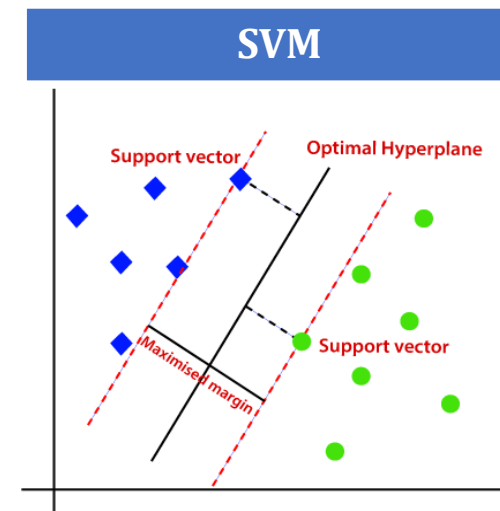


## Floresta Aleatória

### Random Forest Simplified

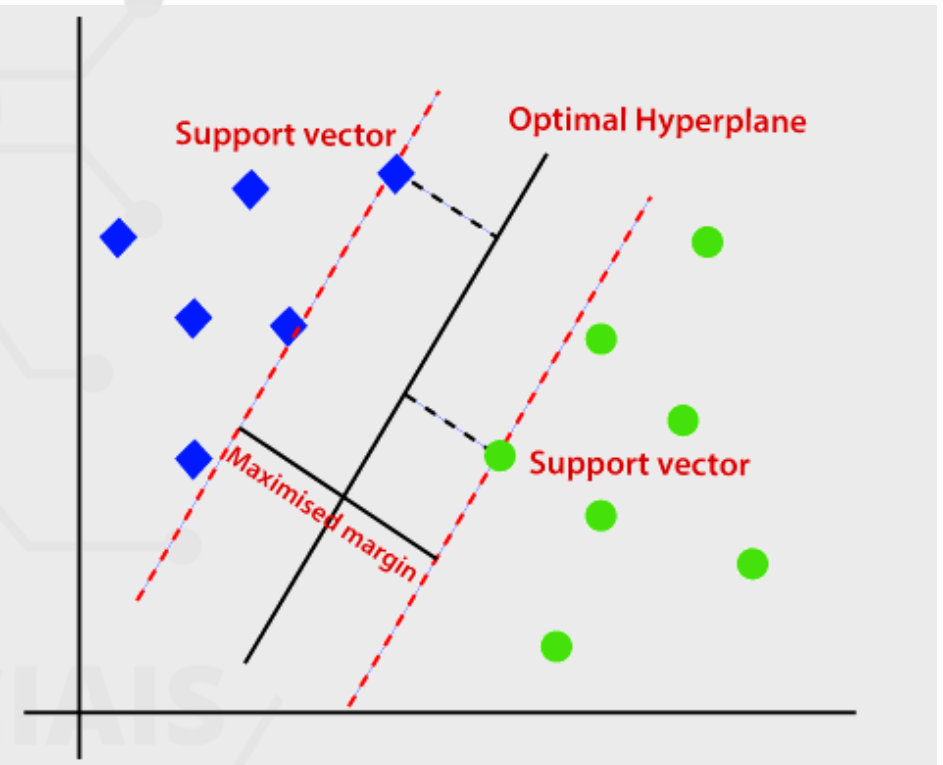


## SVM



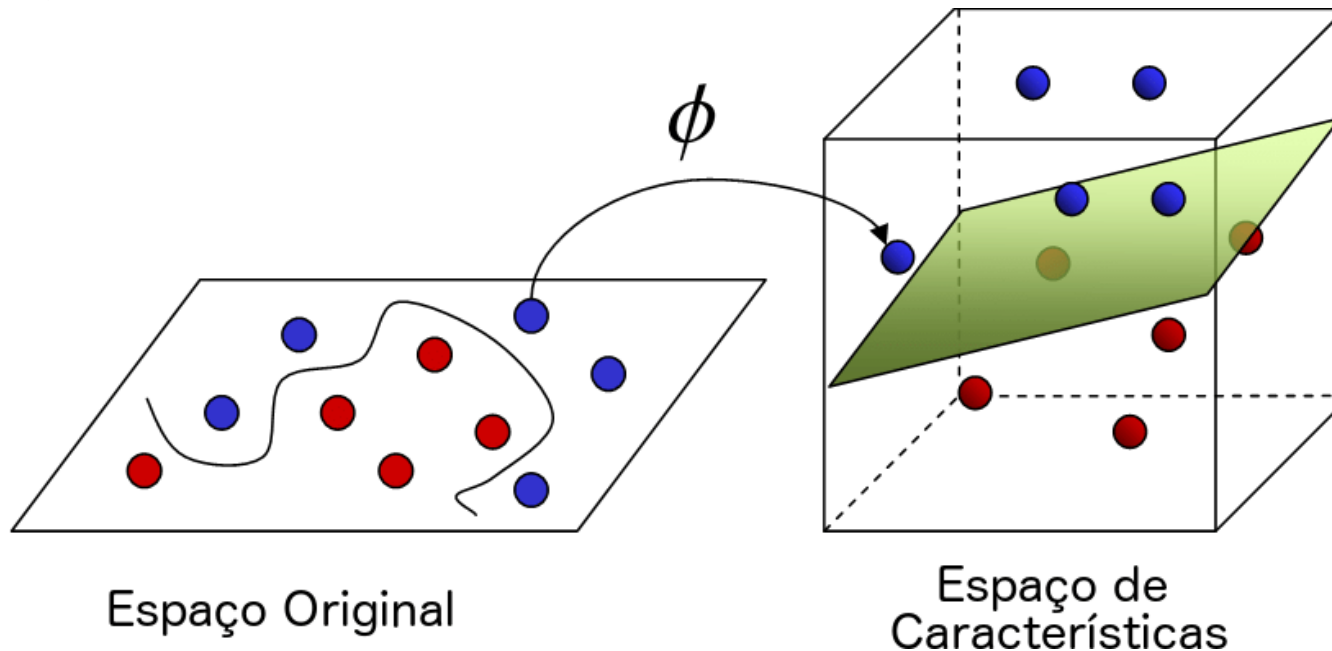
# Support Vector Machine (SVM)

Support Vector Machine é uma fronteira que melhor segrega as duas classes (hiperplano/ linha). Os Vetores de Suporte são as coordenadas da observação individual.



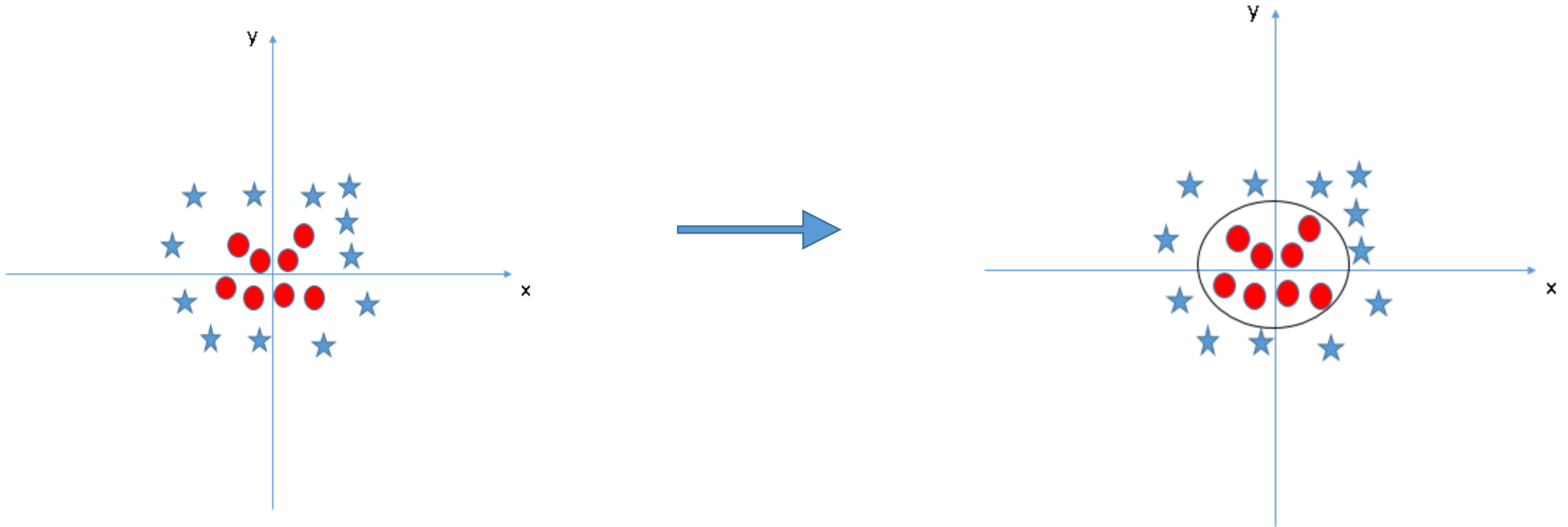
# SVM – Classificação Não-Linear

Num espaço de maior dimensão, espera-se que o problema de classificação se torne linearmente separável. Assim, uma fronteira de decisão linear (hiperplano) pode ser usada para realizar a classificação no espaço  $R^K$ .



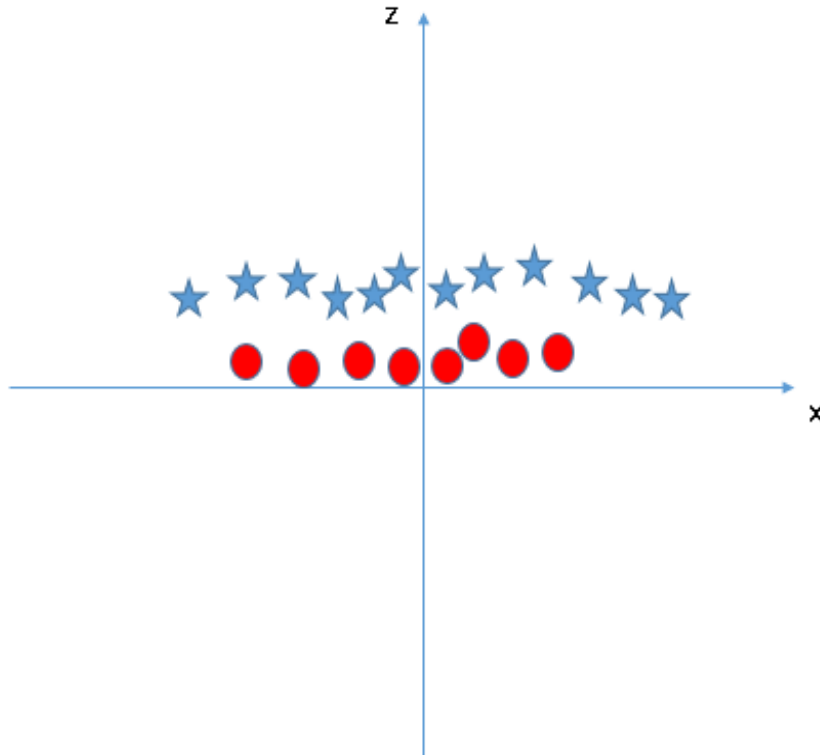
# SVM – Classificação Não-Linear

❑ No cenário abaixo, não podemos ter um hiperplano linear entre as duas classes, então como o SVM classifica essas duas classes?



# SVM – Classificação Não-Linear

- Com um recurso adicional. Aqui, vamos adicionar um novo recurso  $z = x^2 + y^2$ . Agora, vamos plotar os pontos de dados no eixo  $x$  e  $z$ :



No gráfico acima, os pontos a serem considerados são:

- Todos os valores para  $z$  seriam positivos sempre porque  $z$  é a soma quadrática de  $x$  e  $y$ ;
- No gráfico original, os círculos vermelhos aparecem próximos da origem dos eixos  $x$  e  $y$ , levando a um valor menor de  $z$  e estrela relativamente longe do resultado da origem para um valor maior de  $z$ .



# SVM – Classificação Não-Linear

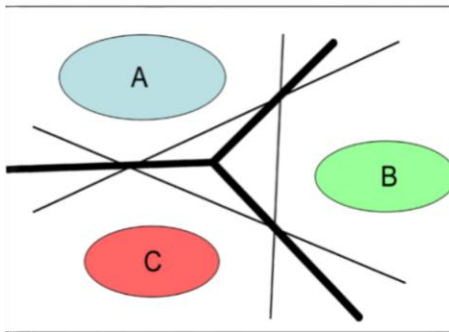
## ➤ Como adicionar esse recurso?

O SVM tem uma técnica chamada **truque** do [kernel](#). Estas são funções que ocupam um espaço de entrada dimensional baixo e o transformam em um espaço dimensional mais alto, convertendo um problema não separável em um problema separável.

**kernel** : Existem várias opções disponíveis com o kernel, como **“linear”**, **“rbf”**, **“poly”** e outros (o valor padrão é “rbf”). Aqui, “rbf” e “poly” são úteis para o hiperplano não linear.

# SVM - Multiclasses

- ❖ Nesse caso, precisamos de múltiplos SVMs binários para construir um classificador multi-classe.

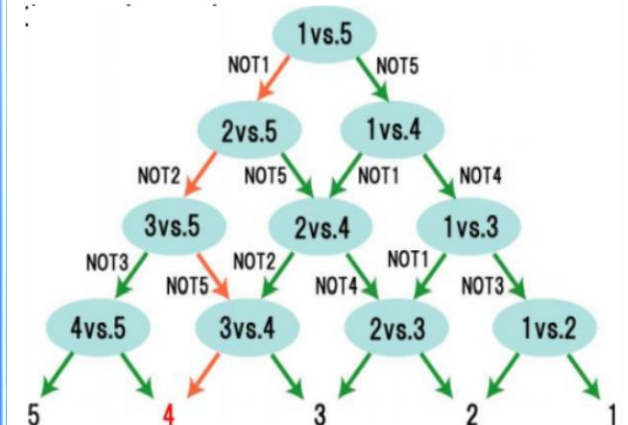


## Decomposição 1-de-n

- $n$  classificadores binários
- Cada classificador identifica uma classe das demais  $(n-1)$  classes restantes

## Decomposição 1-1

- $n*(n-1)/2$  classificadores binários
- Cada classificador classifica uma amostra dentre um par de classes possíveis
- No treinamento, padrões que não pertençam as 2 classes envolvidas são ignorados



# Support Vector Machine (SVM)



## Prós

- Funciona muito bem com margem de separação clara.
- É eficaz em espaços dimensionais elevados.
- É eficaz nos casos em que o número de dimensões é maior que o número de amostras.
- Usa um subconjunto de pontos de treinamento na função de decisão (chamados de vetores de suporte), portanto, também é eficiente em termos de memória.

## Contra

- Não funciona bem quando temos um grande conjunto de dados porque o tempo de treinamento necessário é maior
- Também não funciona muito bem, quando o conjunto de dados tem mais ruído, ou seja, as classes de destino estão sobrepostas
- O SVM não fornece estimativas de probabilidade diretamente, elas são calculadas usando uma valiosa **validação cruzada** de cinco vezes. É o método SVC relacionado da biblioteca *scikit-learn* do Python.

A stylized illustration of a brain with circuit-like lines extending from it, positioned behind a blue ribbon banner.

**Hora da prática**

**ART|EEEEFICIAIS**

## **Aprendizado supervisionado**

### **Classificação**

- Supervisão: As observações no conjunto de treinamento são acompanhadas por “labels” indicando a classe a que elas pertencem.
- Novas ocorrências são classificadas com base no conjunto de treinamento.

## **Aprendizado Não Supervisionado**

### **Clusterização**

- Não existe classe pré-definida para nenhum dos atributos.
- Um conjunto de observações é dado com o propósito de se estabelecer a existência das classes.

# Clustering



Inferências a partir de conjuntos de dados usando apenas vetores de entrada sem se referir a resultados conhecidos.

Conjunto de técnicas de prospecção de dados que visa fazer agrupamentos automáticos de dados segundo o seu grau de semelhança.

**Aprendizado não supervisionado**

**O armazenamento em cluster** é a tarefa de dividir a população ou os pontos de dados em vários grupos.

Os pontos de dados nos mesmos grupos devem ser mais semelhantes a outros pontos de dados no mesmo grupo e diferentes dos pontos de dados em outros grupos.

# Aplicações de clustering

## Marketing

- Caracterização e descoberta de segmentos de clientes para fins de marketing.

## Biologia

- Classificação entre diferentes espécies de plantas e animais.

## Bibliotecas

- Usado para agrupar diferentes livros com base em tópicos e informações.

## Seguro

- Usado para reconhecer os clientes, suas políticas e identificar as fraudes.

## Planejamento da cidade

- Criação de grupos de casas e estudar seus valores com base em sua localização geográfica e outros fatores presentes.

## Estudos de terremotos

- Ao aprender as áreas afetadas pelo terremoto, podemos determinar as zonas perigosas.

# Métodos de Clustering

## Métodos baseados em densidade

- Consideram os aglomerados como a região densa com alguma semelhança e diferente da região densa inferior do espaço.

## Métodos hierárquicos

- Os clusters formados nesse método formam uma estrutura do tipo árvore com base na hierarquia. Novos clusters são formados usando o anteriormente formado.

## Métodos de particionamento

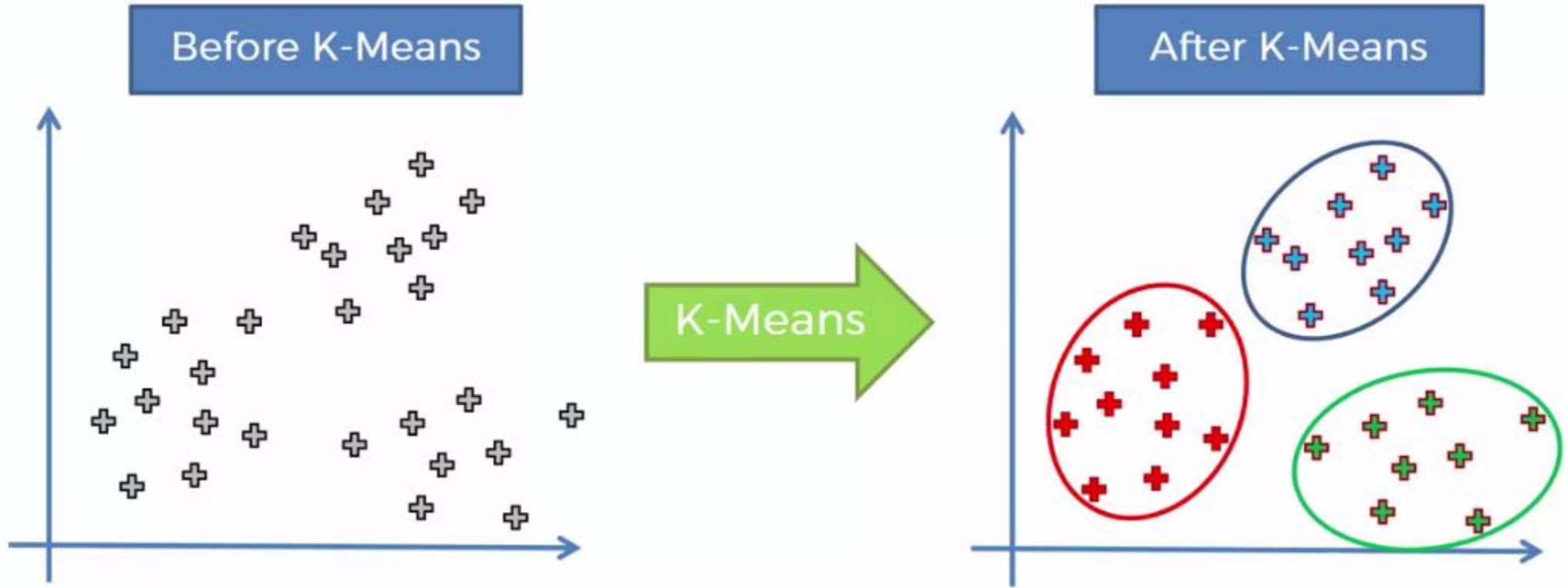
- Esses métodos particionam os objetos em  $k$  clusters e cada partição forma um cluster. Este método é usado para otimizar uma função de similaridade de critério objetivo, como quando a distância é um exemplo de parâmetro principal.

## Métodos baseados em grade

- O espaço de dados é formulado em um número finito de células que formam uma estrutura semelhante a uma grade. Todas as operações de armazenamento em cluster realizadas nessas grades são rápidas e independentes do número de objetos de dados.



# K- Means



# Calculando os centros de agrupamento



The diagram illustrates the three steps of the K-means algorithm. It features a central background image of a brain with circuitry. Three colored boxes represent the steps: a yellow box for Step 1, a green box for Step 2, and a blue box for Step 3. Arrows indicate the flow from Step 1 to Step 2, and from Step 2 to Step 3. A large, faint watermark 'ARTIFICIAIS' is visible across the bottom of the diagram.

- Fora dos pontos a serem agrupados, atribua aleatoriamente  $n$  pontos como o centro do cluster, onde  $n$  é o número de clusters necessários.

**Etapa 1**

**Etapa 2**

- encontre a distância entre todos os pontos com os pontos escolhidos como centros aleatórios.

- atribua os pontos ao cluster ao qual ele está mais próximo.

**Etapa 3**

A stylized illustration of a brain with circuit-like lines extending from it, positioned behind a blue ribbon banner.

**Hora da prática**

**ART|EEEEFICIAIS**

A stylized graphic of a brain, split vertically. The left half is purple and the right half is blue. Both halves have a circuit-like pattern of lines and dots. The word "OBRIGADA!" is written in large, bold, black capital letters across the center of the brain.

# OBRIGADA!

Repositório GitHub:

<https://github.com/Skyzenho/ArtIEEEficiais>

**ARTIEEEFICIAIS**