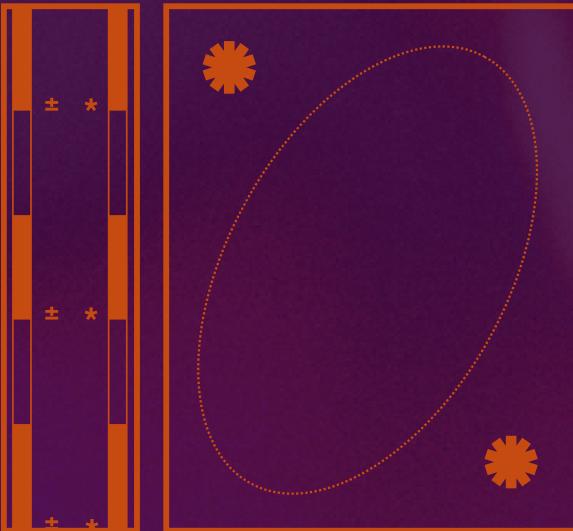


APRENDIZADO DE MÁQUINA



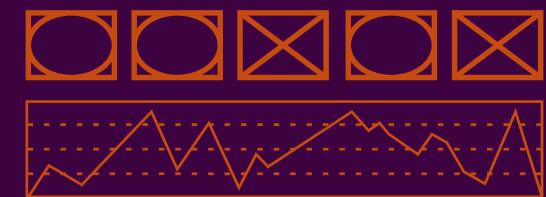
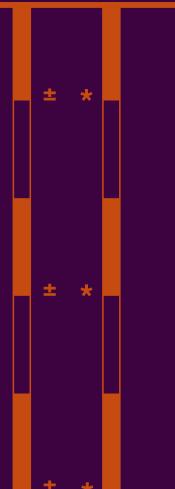
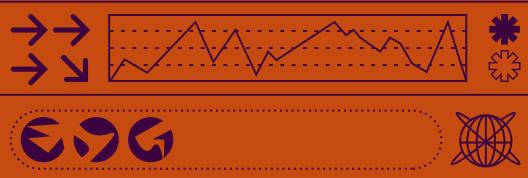
DIA 3



TÓPICOS



Tópicos explorados na Aula:



- » Árvores de Decisão
- » Métricas para Seleção de Atributos
- » Random Forest
- » Clustering
- » Aprendizado por Reforço
- » Seleção de Modelos

AGENDA AGENDA AGENDA AGENDA AGENDA AGENDA

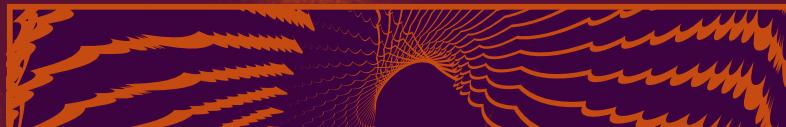


EDG



ÁRVORES DE DECISÃO

- O que são Árvores de Decisão?
- Como construir e utilizar?



ÁRVORES DE DECISÃO

- **Nó Raiz:** é o ponto de partida, com o conjunto de dados inteiro e a primeira pergunta;
- **Nós internos:** dividem os dados através de perguntas (testes) sobre um atributo.
- **Ramos:** as possíveis saídas da pergunta.
- **Nós folhas:** contêm a predição final (classe ou valor) para o seu subconjunto de dados.

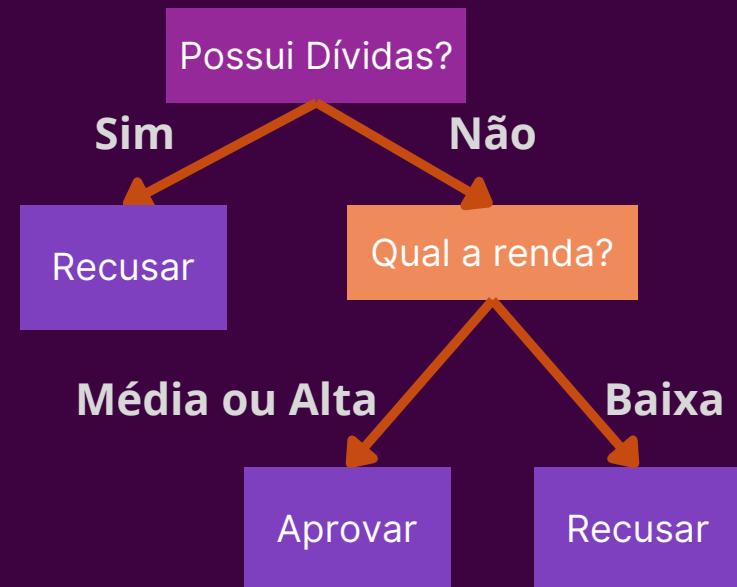


ÁRVORES DE DECISÃO

COMO SÃO USADAS?

- Há uma observação cujo rótulo queremos descobrir;
- Os valores de seus atributos são testados nos nós da Árvore;
- Em algum momento, o caminho chegará a uma folha → rótulo que a Árvore prediz para o dado.

Aprovação de Empréstimo



ÁRVORES DE DECISÃO

APRENDIZADO SUPERVISIONADO

- Árvores de Decisão são modelos de aprendizado supervisionado;
- Sua construção é feita a partir de dados de treinamento rotulados.

CONSTRUÇÃO - INDUÇÃO

- O objetivo do modelo é dividir os dados em subconjuntos, onde cada elemento tem o mesmo rótulo;
- Essa divisão segue uma hierarquia de atributos (nó raiz trata o atributo mais importante, que divide os dados da forma mais “pura” possível, por exemplo).

ÁRVORES DE DECISÃO

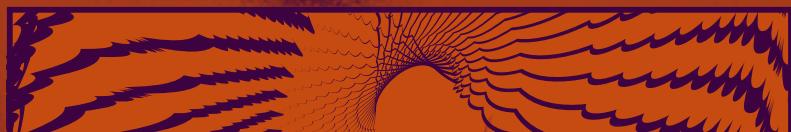
VANTAGENS

- A própria Árvore decide sobre quais atributos irá operar → possivelmente reduzindo a dimensionalidade dos dados;
- Conhecimento construído é acessível → relativamente simples de entender;
- Construção da Árvore não exige determinação e calibração dos parâmetros;
- O treinamento e o uso da árvore para predição é rápido.



MÉTRICAS PARA SELEÇÃO DE ATRIBUTOS

- Quais são?
- Como são calculadas?



SELEÇÃO DE ATRIBUTOS

Para decidir qual é o melhor atributo para fazer a divisão do conjunto de dados, são utilizadas métricas que calculam um valor para cada atributo, que representam a “pureza” ou "impureza" de um conjunto de dados.

Essa seleção dos melhores atributos permite a criação do melhor modelo, que faz as classificações mais exatas.

As mais comuns são:

- **Entropia e ganho de informação;**
- **Índice de Gini.**

ENTROPIA

- A entropia é uma métrica que quantifica o grau de incerteza ou desordem em um conjunto de dados.
- O objetivo das Árvores de Decisão é reduzir a entropia dos conjuntos a cada divisão.

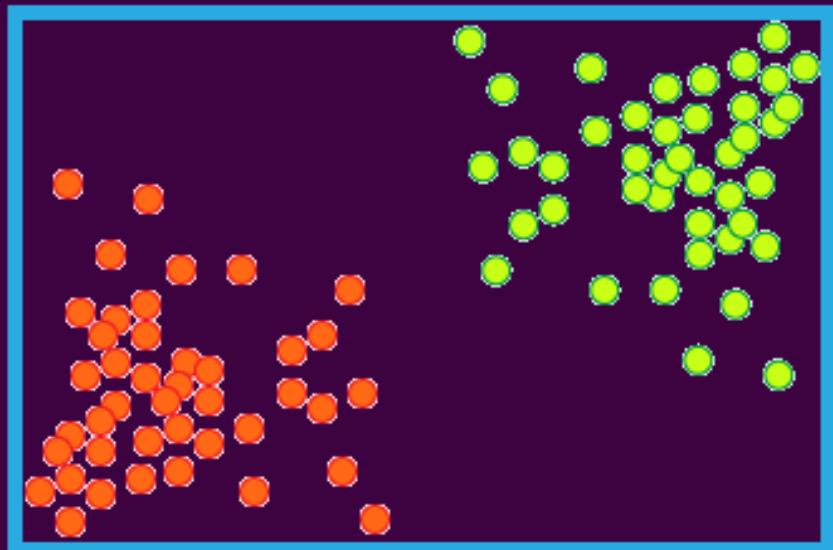
$$\text{Entropia}(S) = - \sum_{i=1}^c p_i \log_2(p_i)$$

c: o número de classes.

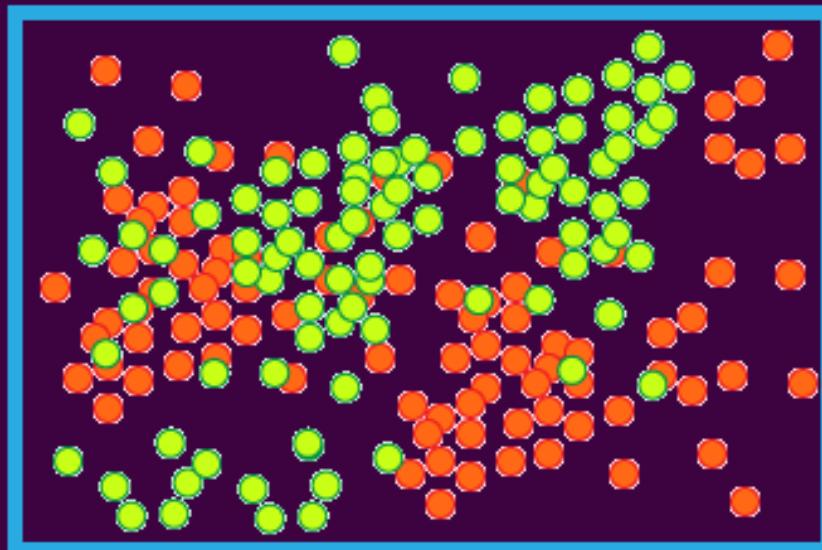
Pi: proporção de observações da classe i no conjunto de dados S.

ENTROPIA

GRAU DE INCERTEZA NOS DADOS



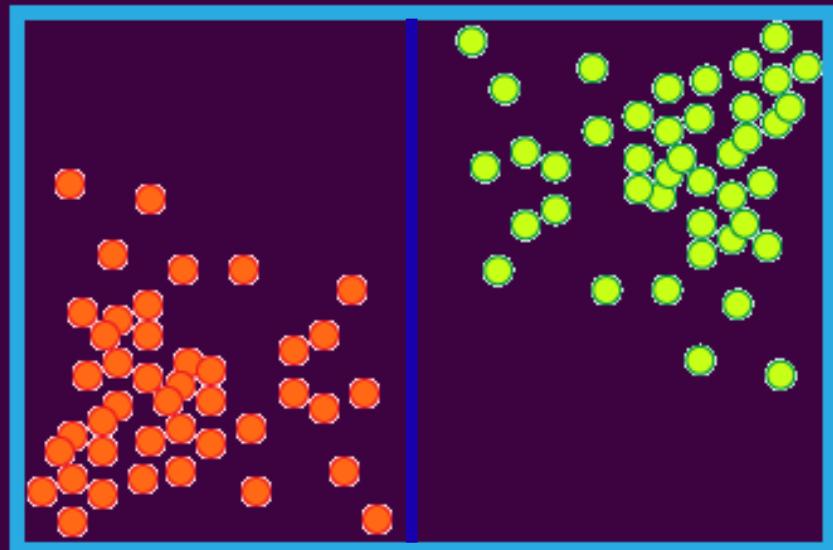
CONJUNTO DE DADOS 1



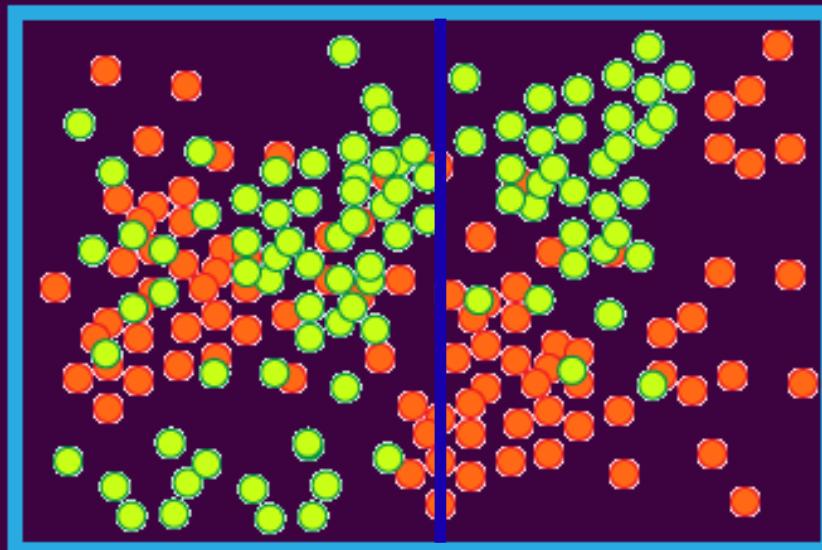
CONJUNTO DE DADOS 2

ENTROPIA

GRAU DE INCERTEZA NOS DADOS



ENTROPIA = 0
SUBCONJUNTOS PUROS



ENTROPIA ALTA = ~1
SUBCONJUNTOS IMPUROS

ENTROPIA

BAIXA ENTROPIA:

- Significa que um grupo de dados é muito homogêneo
- Pouca incerteza e alta previsibilidade
- Ex: um grupo de clientes onde 95% compraram o Produto A e apenas 5% compraram o Produto B. Há uma alta certeza sobre a preferência.

ALTA ENTROPIA:

- Significa que um grupo de dados é muito heterogêneo
- Resulta em muita incerteza ou baixa previsibilidade
- Ex: um grupo de clientes onde 50% compraram o Produto A e 50% o Produto B. Há muita incerteza sobre a preferência individual.

GANHO DE INFORMAÇÃO

- **O Ganho de Informação é uma métrica que mede a redução da entropia após a divisão de um conjunto de dados com base em um determinado atributo.**
- **Ele serve como critério para a seleção do melhor atributo para dividir os dados em um nó da árvore.**
- **O melhor atributo é o que possui o maior valor de ganho de informação, pois ele resulta nos subconjuntos mais homogêneos.**

$$\text{Ganho}(S, A) = \text{Entropia}(S) - \sum_{v \in \text{Valores}(A)} \frac{|S_v|}{|S|} \times \text{Entropia}(S_v)$$

ÍNDICE DE GINI

- O Índice de Gini é outra métrica, alternativa à entropia e ao ganho de informação, que mede a pureza de um nó.
- Um valor de 0 indica que o conjunto é completamente puro.
- Um valor de 0,5 (para um problema com duas classes) indica a máxima impureza, com uma distribuição uniforme das classe.

$$\text{Gini}(S) = 1 - \sum_{i=1}^c p_i^2$$

c: o número de classes.

Pi: proporção de observações da classe i no conjunto de dados S.

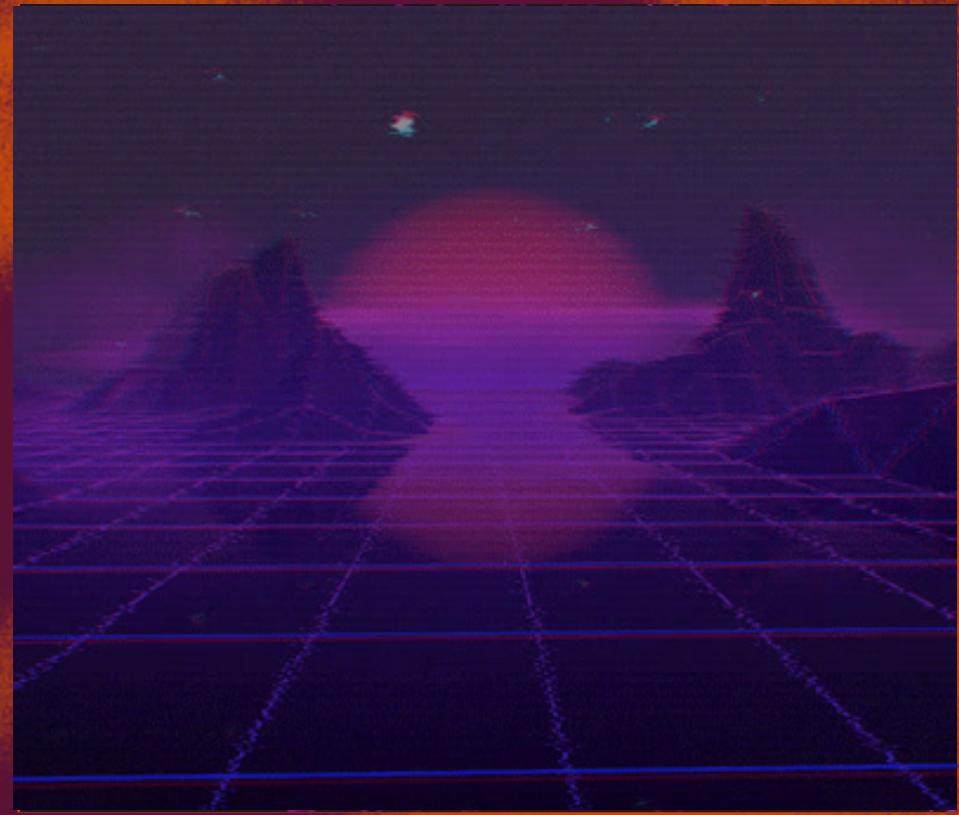


E.D.G



OTIMIZAÇÕES

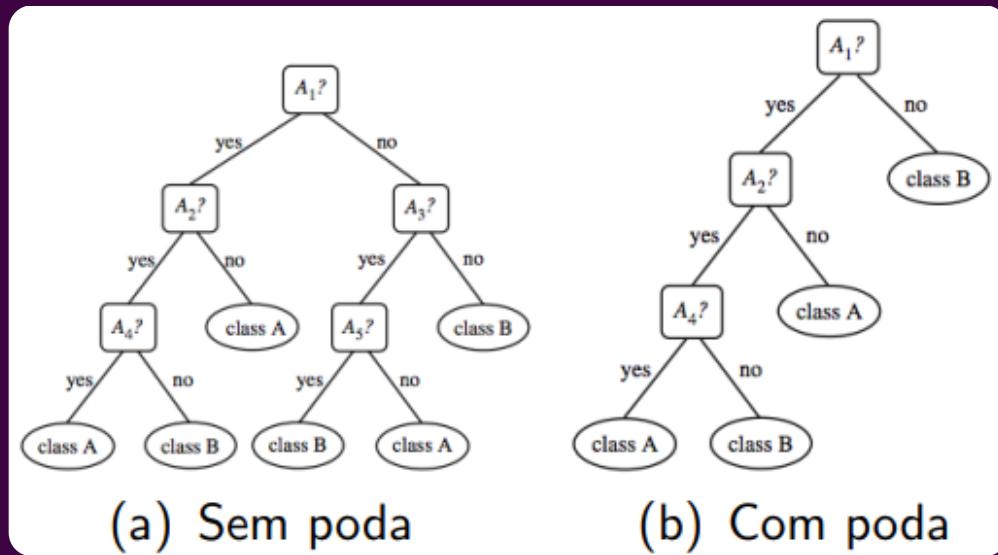
- Poda (pruning)
- Hiperparâmetros



ÁRVORES DE DECISÃO

PRUNING - PODA

É uma técnica utilizada para evitar o Overfitting do modelo.



ÁRVORES DE DECISÃO

PRUNING - PODA

- Duas abordagens:
 - Pré-poda → Parar o crescimento da árvore antes que ela se torne muito complexa, utilizando critérios de parada como:
 - Profundidade máxima da árvore;
 - Ganho de informação mínimo.
 - Pós-poda → elimina ramos após a construção completa da Árvore. Isso é feito avaliando o impacto da remoção de um ramo no desempenho da árvore (acurácia, F1-score, erro).

HIPERPARÂMETROS

Principais hiperparâmetros do modelo de Árvores de Decisão:

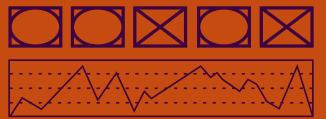
- **Criterion:** a função que mede a qualidade das divisões da árvore.
Valores: Gini, Entropy, Log loss.
- **Max_depth:** a profundidade máxima da árvore
- **Min_samples_split:** O número mínimo de amostras que um nó deve ter para poder ser dividido em novos nós.
- **Min_impurity_decrease:** um nó só será dividido se resultar em uma redução da impureza que seja maior ou igual ao valor definido.

DOCUMENTAÇÃO



EXEMPLO PYTHON

co



RANDOM FOREST

- O que é Random Forest?
- Qual a diferença para Árvores de Decisão?
- Há vantagens?

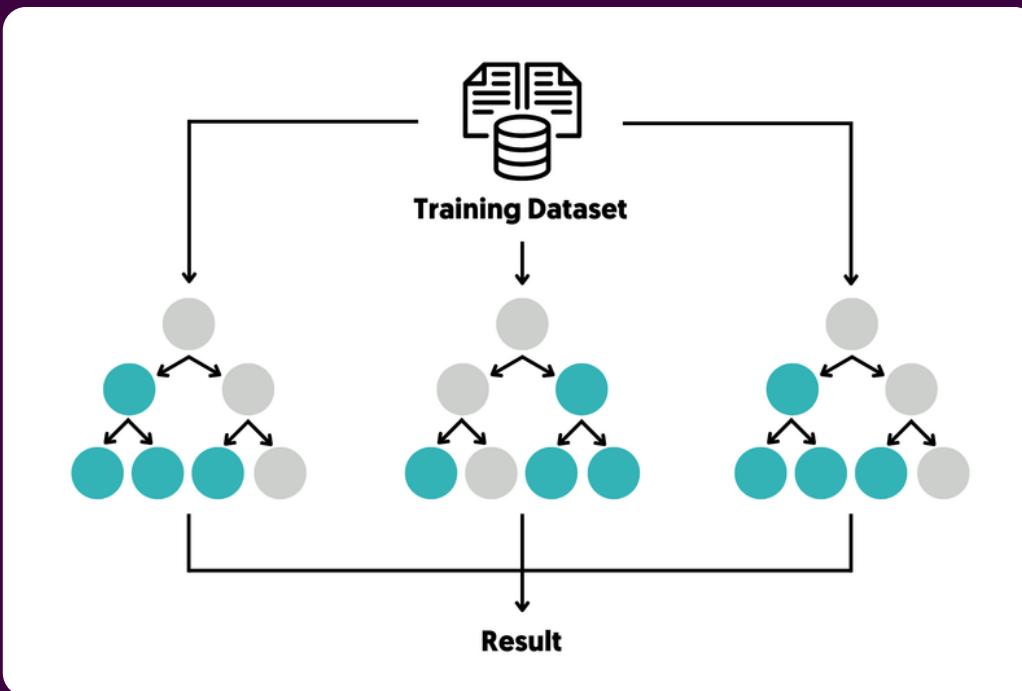


本工事は日本で最も古い木造建築物で、その歴史は約1500年にも及ぶ。この木造建築は、自然災害に対する耐久性と、伝統的な建築技術の継承を示す重要な文化財である。また、この木造建築は、日本の伝統的な建築様式である「和風」の代表的な例として、世界中の建築愛好家に高く評価されている。



RANDOM FOREST

Conjunto de Árvores de Decisão.



RANDOM FOREST

CONSTRUÇÃO

- **Bootstrap:** São criadas amostras com reposição dos dados de treino;
- Os dados que não são selecionados para uma amostra específica são usados para validação;
- Cada amostra é utilizada para a construção de uma Árvore;
- Cada Árvore utiliza um subconjunto de atributos aleatórios menor que o original.

UTILIZAÇÃO

- Cada dado de teste é passado a todas as Árvores;
- O rótulo que for majoritariamente “votado” pelas árvores é dado como a resposta final do modelo.

RANDOM FOREST

QUANDO VALE A PENA USAR

- **Cada Árvore do conjunto deve ser simples:**
 - **Tempo de treinamento viável;**
 - **Ainda assim, com algum padrão de qualidade.**
- **Cada Árvore deve errar em aspectos e dados diferentes (quando erram):**
 - **Diversidade entre as Árvores.**

RANDOM FOREST

VANTAGENS VS ÁRVORE DE DECISÃO

- **Performance muito superior e mais robusta que uma única árvore;**
- **Menos problemas com overfitting;**

DESVANTAGENS VS ÁRVORE DE DECISÃO

- **Demande mais poder computacional e/ou tempo para treinamento e para realização das previsões.**

HIPERPARÂMETROS

Principais hiperparâmetros do modelo de Árvores de Decisão:

- **Criterion:** a função que mede a qualidade das divisões da árvore.
Valores: Gini, Entropy, Log loss.
- **Max_depth:** a profundidade máxima da árvore
- **Min_samples_split:** O número mínimo de amostras que um nó deve ter para poder ser dividido em novos nós.
- **Min_impurity_decrease:** um nó só será dividido se resultar em uma redução da impureza que seja maior ou igual ao valor definido.

HIPERPARÂMETROS

Principais hiperparâmetros do modelo de Random Forest:

- **n_estimators:** O número de árvores de decisão que serão criadas na floresta.
- **max_features:** O número máximo de características (atributos) que é considerada na seleção da melhor feature para cada divisão de nó da árvore.
obs: valor padrão é a raiz quadrada do número total de features.

DOCUMENTAÇÃO



GRIDSEARCHCV

DEFININDO HIPERPARÂMETROS

O Grid-Search busca a melhor combinação de hiperparâmetros para que o modelo tenha o melhor desempenho possível.

Funcionamento do Grid-Search:

- 1. Definir uma lista de valores para testar para cada hiperparâmetro;**
- 2. Ao executar o Grid-Search ele irá treinar (com cross-validation) e avaliar um modelo Random Forest para cada combinação possível de hiperparâmetros;**
- 3. Por fim, ele informa a melhor combinação de hiperparâmetros para o modelo.**

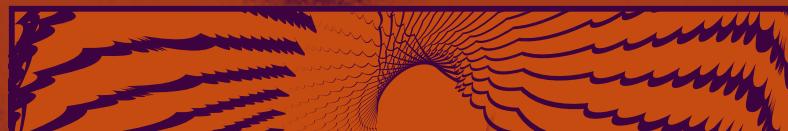
EXEMPLO PYTHON

co



CLUSTERING

- O que é?
- Qual sua finalidade?
- K-Means



CLUSTERING

○ QUE É?

É a tarefa de agrupar um conjunto de objetos de tal forma que objetos no mesmo grupo (chamado de cluster) são mais similares entre si do que com aqueles em outros grupos.

Diferente do aprendizado supervisionado (como classificação e regressão), na clusterização nós não temos um "rótulo" ou uma "resposta correta" para os dados. O algoritmo explora a estrutura dos dados por conta própria para encontrar padrões ou agrupamentos naturais.

CLUSTERING

UTILIDADES

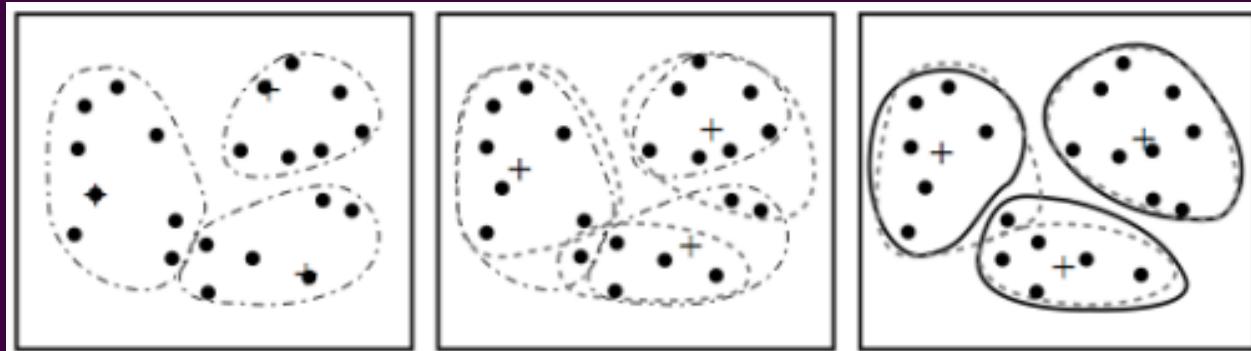
- **Marketing e negócios: agrupar clientes com base em seu comportamento de compra.**
- **Biologia e genética: agrupar genes com padrões de expressão semelhantes.**
- **Medicina: identificar grupos de pacientes com sintomas ou respostas a tratamentos similares.**
- **Processamento de Imagens: agrupar pixels de cores semelhantes para identificar objetos**

K-MEANS

“EXTRAIR” K GRUPOS DOS DADOS

- **K:**

- É um parâmetro definido por nós no modelo;
- Corresponde à quantidade de protótipos que serão criados.



K-MEANS

COMO ESCOLHER K?

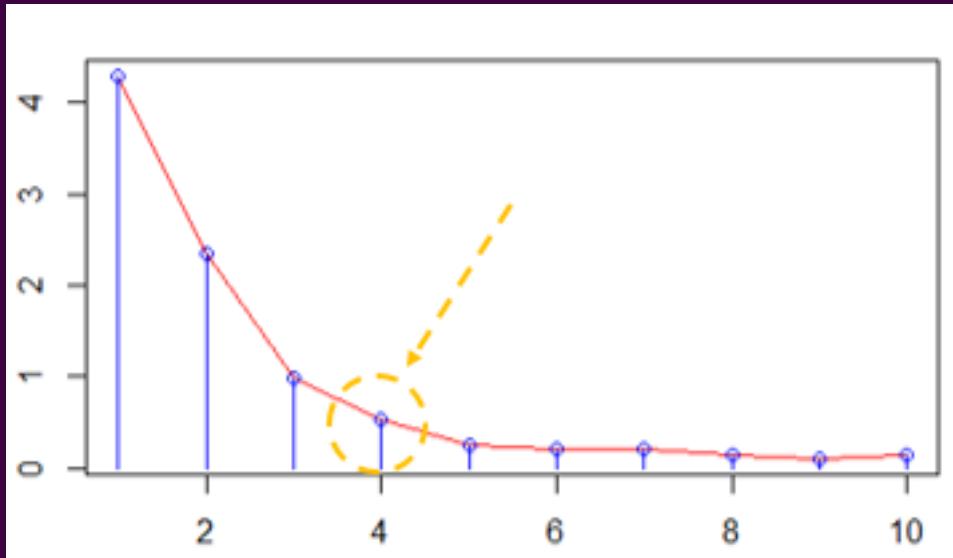
- **Método do Cotovelo:**

- **Mede a soma dos quadrados das distâncias dentro de cada cluster;**
- **Permite comparar os efeitos de cada quantidade diferente de clusters criados (K).**

K-MEANS

COMO ESCOLHER K?

- **Método do Cotovelo:**



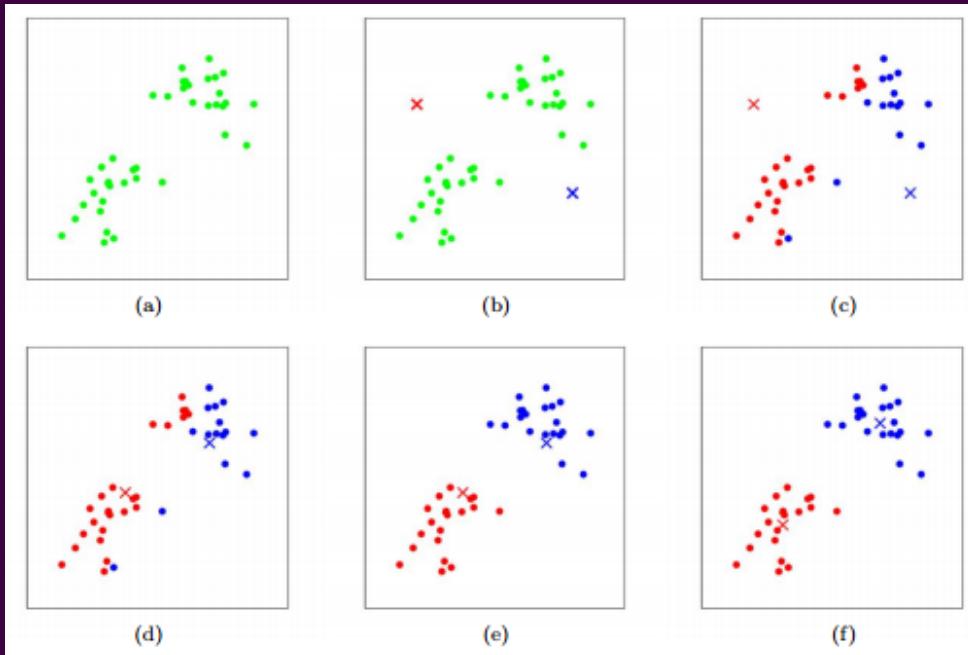
K-MEANS

FUNCIONAMENTO

- São criados aleatoriamente no espaço vetorial K protótipos (vetores);
- Para cada dado, é calculada a distância até cada um dos protótipos, a fim de encontrar aquele que está mais próximo;
- O dado passa a fazer parte do grupo do protótipo mais próximo (mais similar);
- A partir dos dados que agora estão em seu grupo, o protótipo se ajusta para representá-los melhor (se posiciona na média dos dados);
- O processo segue até que convirja.

K-MEANS

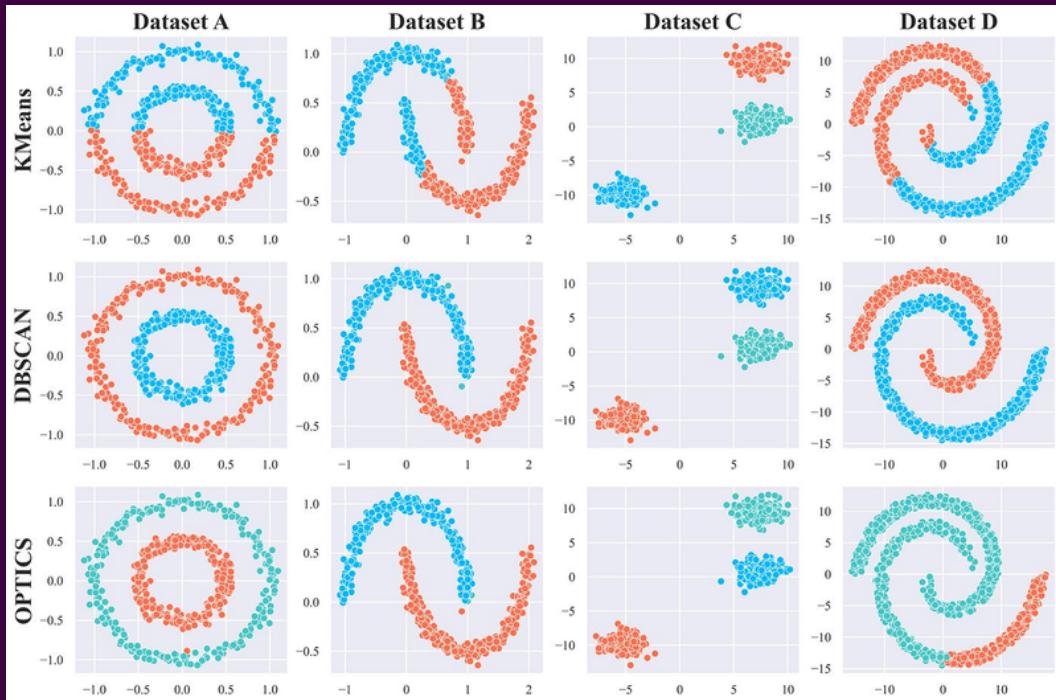
FUNCIONAMENTO



K-MEANS

RESOLVE QUALQUER PROBLEMA?

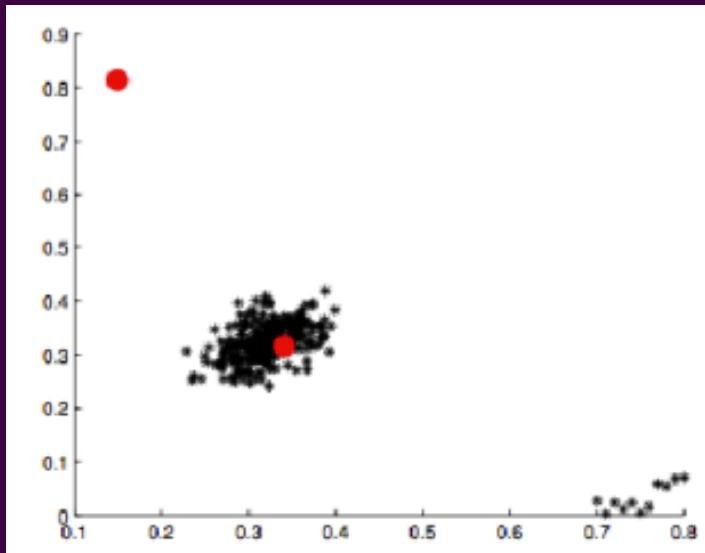
NÃO!!



K-MEANS

OTIMIZAÇÕES

- Inicialização dos protótipos no espaço vetorial:



K-MEANS

OTIMIZAÇÕES

- Inicialização dos protótipos no espaço vetorial:
 - Tradicionalmente, como vimos, é aleatória;
 - Dependendo da configuração dos dados e das posições iniciais dos protótipos, podemos chegar a resultados indesejados.

VALIDAÇÃO DE CLUSTERS

ÍNDICES EXTERNOS

- Os índices externos comparam os agrupamentos obtidos com uma classificação real pré-existente (rótulos verdadeiros).
- Servem para avaliar o quanto bem o algoritmo conseguiu reproduzir uma estrutura conhecida.
- Eles verificam se os elementos que pertencem à mesma classe real foram agrupados corretamente, e se os de classes diferentes foram separados.

VALIDAÇÃO DE CLUSTERS

ÍNDICES INTERNOS

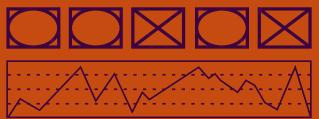
Os índices internos avaliam a qualidade dos agrupamentos considerando apenas os próprios dados, sem necessidade de rótulos verdadeiros.

Índice Silhouette:

$$S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

a(i): é a distância média do dado i a todos os demais dados do seu grupo

b(i): é a distância média mínima do dado i a todos os dados de cada um dos demais grupos (excluindo o seu)

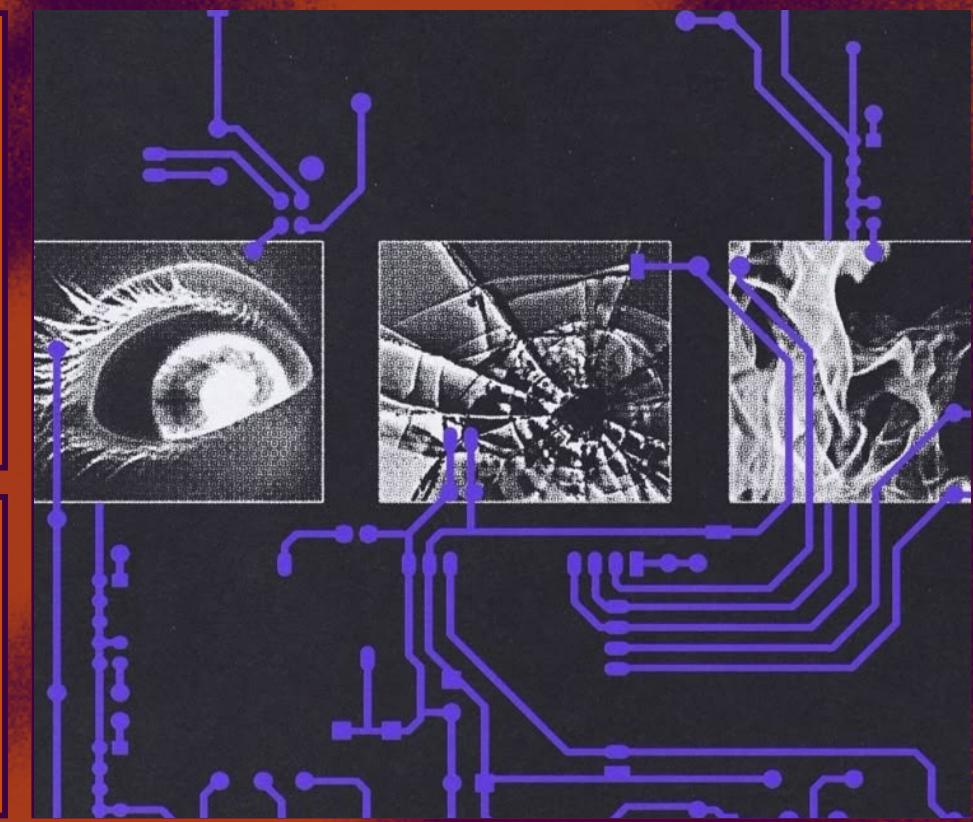


EDG



APRENDIZADO POR REFORÇO

- O que é?
- Como funciona?
- Onde é usado?



REFORÇO

APRENDER COM A EXPERIÊNCIA

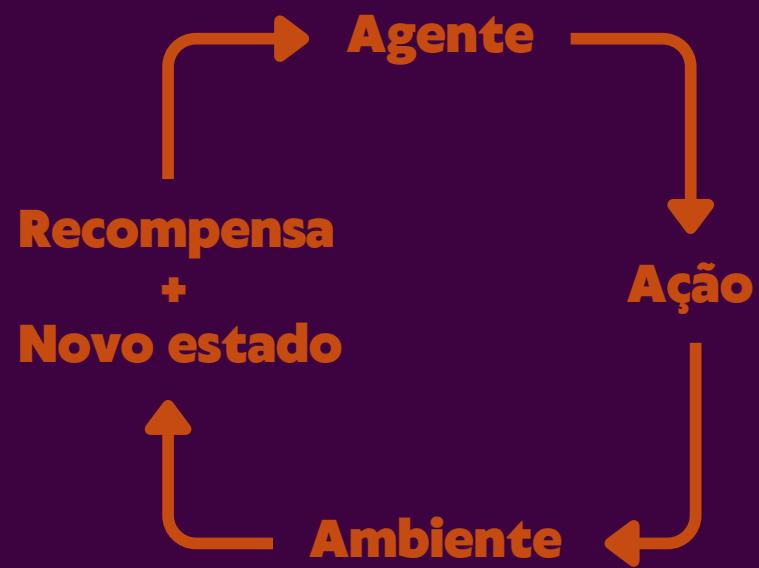
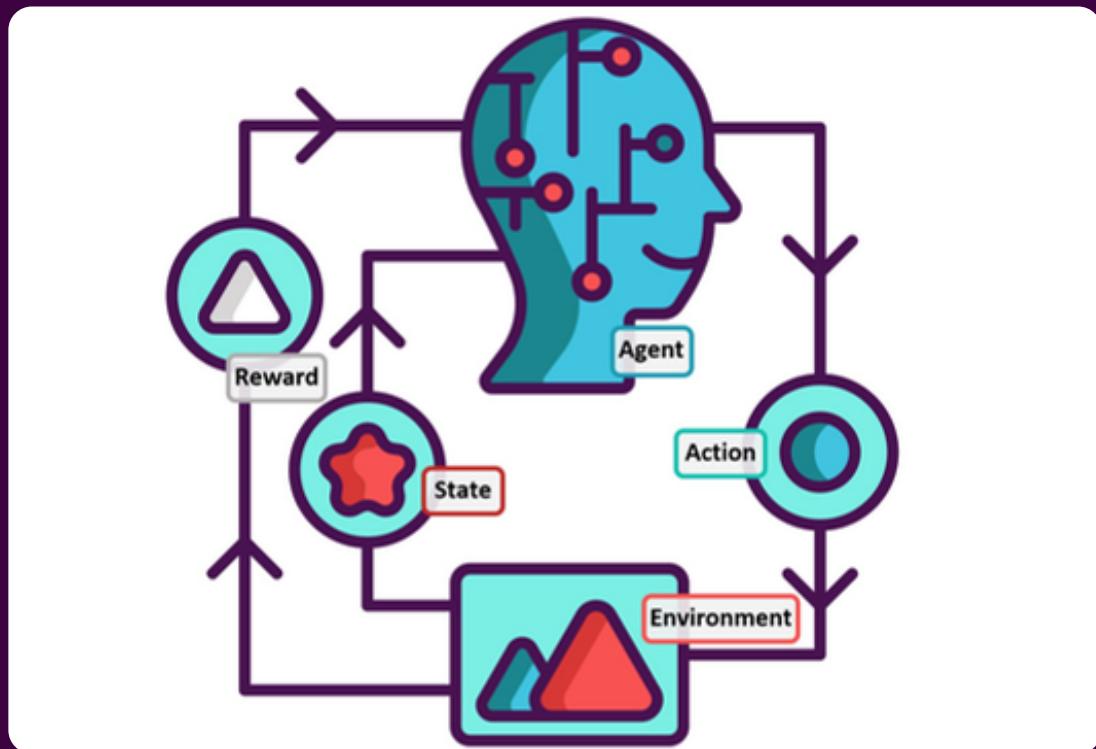
- **Agente:**

- Um “Robô” que precisa aprender algo
- Aprende interagindo com o ambiente
- Quando se sai bem → Recebe uma recompensa
- Quando se sai mal → Não ganha nada ou é punido



É como treinar um cachorro! :D

CICLO DE APRENDIZADO



COMPONENTES

AS PARTES DE UM MODELO DE APRENDIZADO POR REFORÇO

- **Agente:** aquele que aprende e toma as decisões;
- **Ação:** O que o agente pode fazer para mudar o estado;
- **Ambiente:** é o "mundo" onde o agente vive e interage;
- **Estado:** a situação atual do ambiente que o agente observa;
- **Recompensa:** o "feedback" do ambiente, dizendo se a ação foi boa ou ruim. O objetivo é conseguir o máximo de recompensas!

APLICAÇÕES

- Jogos
- Robótica
- Carros autônomos
- Recomendação de conteúdos
- Outros...



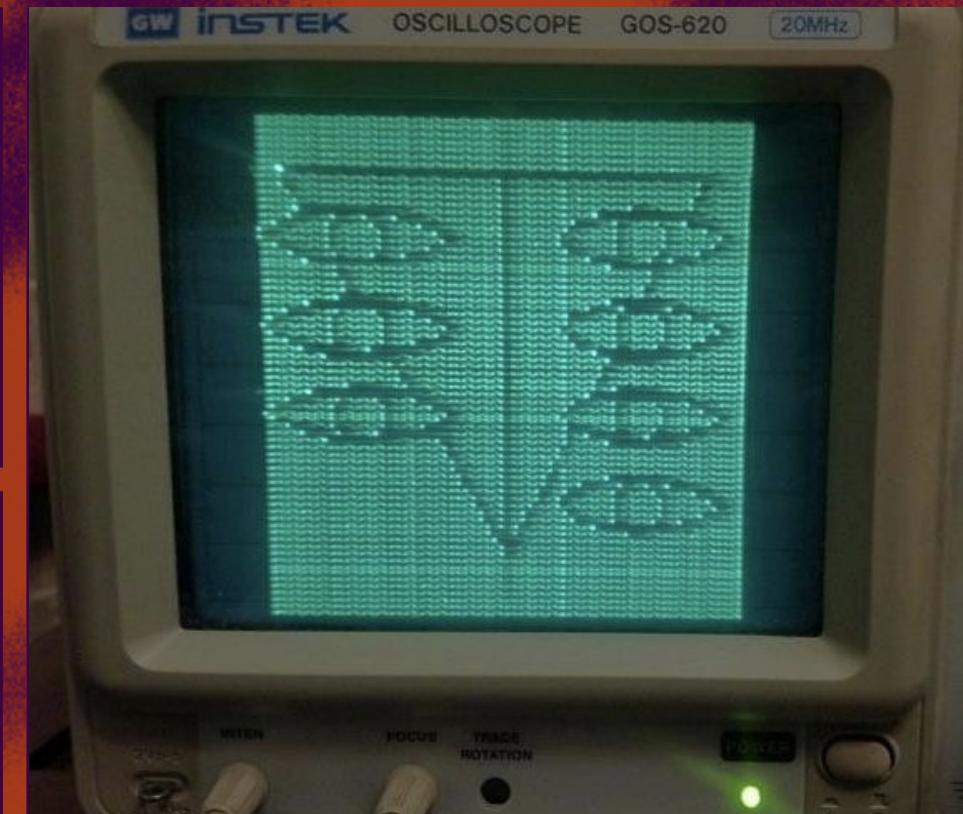


EDG



SELEÇÃO DE MODELOS

- Como saber qual modelo devo usar?
- Quais são as possibilidades?



SELEÇÃO DE MODELOS

CRITÉRIOS A CONSIDERAR

- **Variável resposta do modelo:**
 - **Contínua ou categórica?**
 - **Multiclasse ou binária?**
- **Formato dos dados:**
 - **Imagen, texto, matriz, etc?**
 - **Há rótulos?**
 - **Muitos ou poucos?**
- **Recursos disponíveis:**
 - **Em termos de hardware**
 - **Em termos de tempo**
- **Objetivo do estudo:**
 - **Agrupar?**
 - **Predizer?**
 - **Apontar padrões?**

SELEÇÃO DE MODELOS

ALGUNS CASOS

- **Identificação de animais**
 - **Formato dos dados → imagens**
 - **Objetivo do estudo → criar modelo que informe qual é o animal**
 - **Sugestão → CNN**
- **Segmentação de campanha de marketing**
 - **Formato dos dados → características dos clientes da marca**
 - **Objetivo do estudo → clusterizar os clientes, para criar campanhas direcionadas**
 - **Sugestão → KMeans ou rede SOM**

SELEÇÃO DE MODELOS

ALGUNS CASOS

- **Estudo do mercado de ações**
 - **Formato dos dados → séries históricas**
 - **Objetivo do estudo → sugerir a compra ou venda de ações**
 - **Sugestão → MLP**
- **Entre muitos outros modelos e possibilidades.**