4ª Live

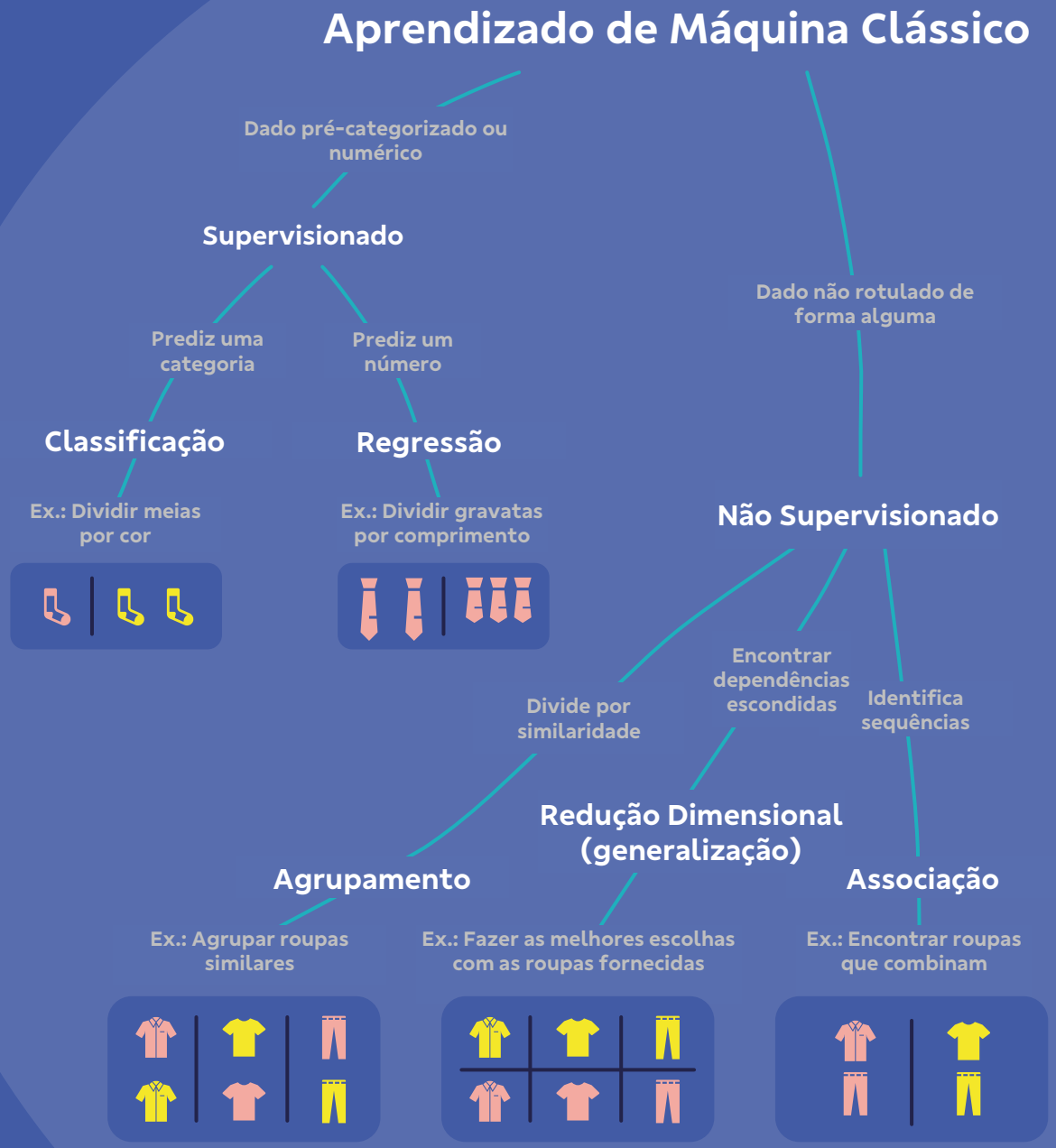
Inteligência Artificial & Ciência de Dados

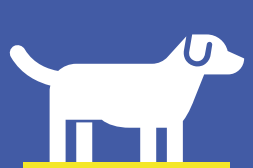
Aprendizado supervisionado

UniBB

Definição

Aprendizado supervisionado é uma abordagem de aprendizado de máquina (machine learning) que utiliza conjuntos de dados **rotulados** para treinar algoritmos. Esses dados rotulados incluem entradas e saídas corretas, permitindo que o modelo aprenda a classificar dados ou prever resultados com precisão.





O que é rótulo?

Sinônimos:
Target, alvo, variável dependente, variável de saída, label, variável resposta.

Em aprendizado supervisionado, **um rótulo é a resposta ou a saída** correta **associada a cada exemplo no conjunto de dados de treinamento.**

Por exemplo, se você está treinando um modelo para reconhecer imagens de gatos e cachorros, cada imagem no conjunto de dados virá com um rótulo indicando se a imagem é de um gato ou de um cachorro.

Os rótulos são essenciais **porque permitem que o algoritmo aprenda a fazer previsões precisas**. Durante o treinamento, o modelo ajusta seus parâmetros para minimizar a diferença entre suas previsões e os rótulos corretos.

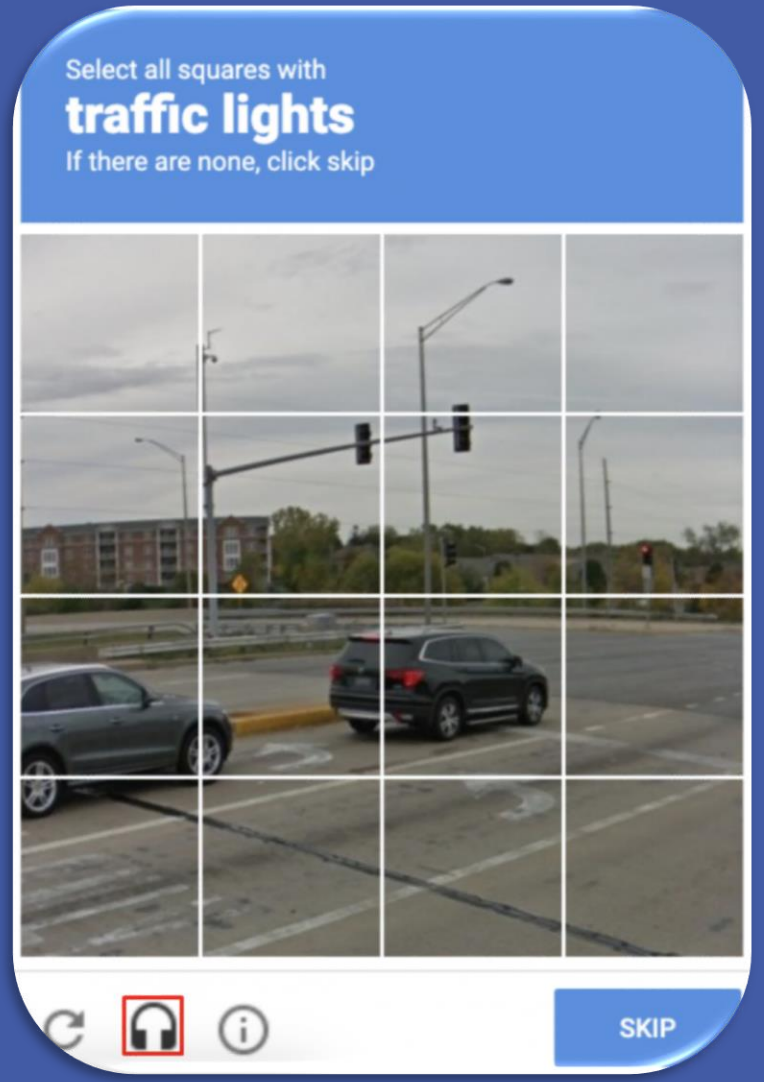
Desafio

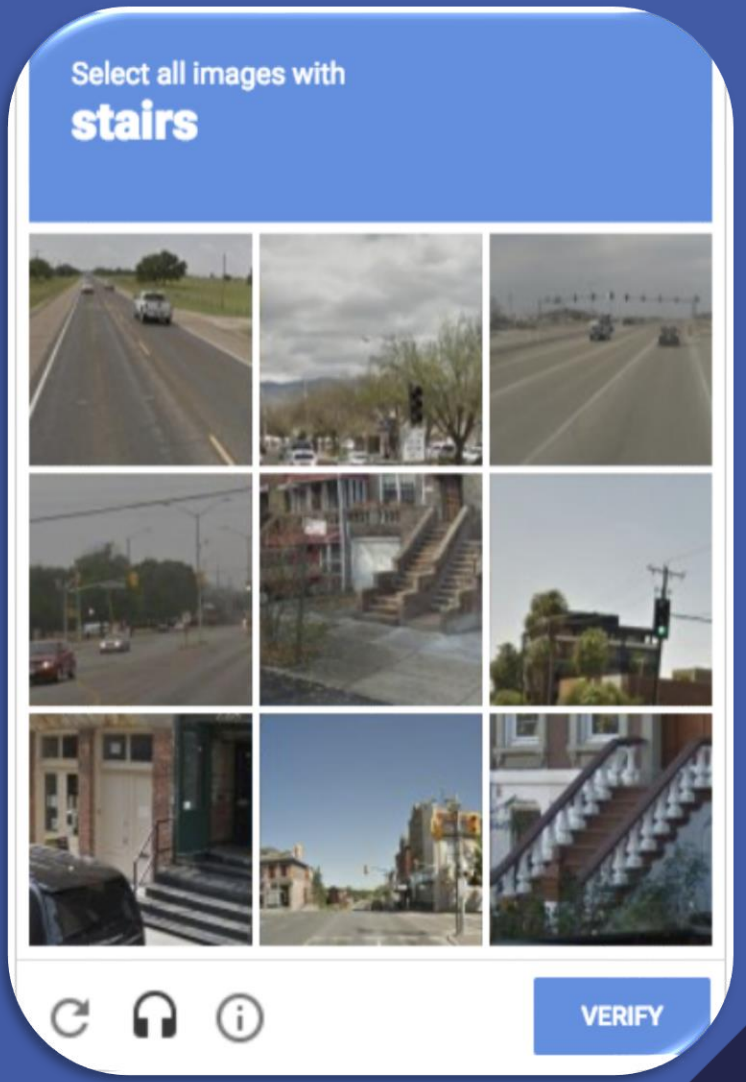
Isso é um cachorro?

Isso é um cachorro?

ACADEMIA BB

Rotulando



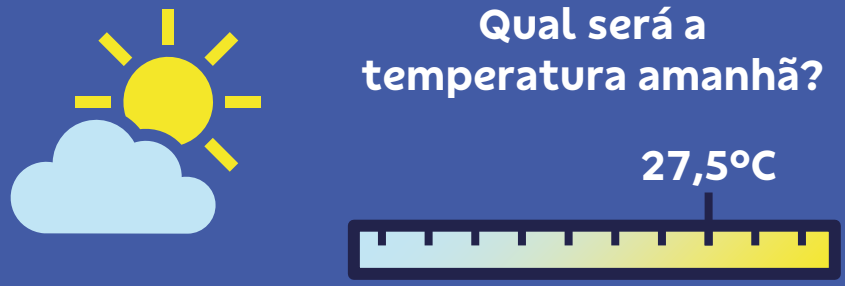


Tarefas supervisionadas

Existem dois tipos principais de problemas que o aprendizado supervisionado pode resolver:

Classificação

envolve treinar um modelo para categorizar dados em classes predefinidas.



Regressão

envolve prever um valor contínuo com base em dados de entrada

Exemplos

Classificação

- Classificação de e-mails (spam)
- Detecção de fraude em transações
- Previsão de churn (cancelamento)
- Análise de sentimentos

Regressão

- Probabilidade de risco de crédito
- Previsão do preço de um imóvel
- Estimativa das vendas futuras de um produto
- Previsão do consumo de energia

Exemplo - Classificação

Vamos considerar um exemplo prático de classificação para **prever o churn de clientes**, ou seja, identificar quais clientes estão propensos a deixar um serviço ou cancelar uma assinatura. Aqui está um passo a passo detalhado:

Variável target

Cliente	Idade	Gênero	Data da última compra	Quantidade de compras em 2024	Fez reclamação no SAC?	Cancelou o serviço?
Bruna	33	F	12/04/2024	2	SIM	SIM
Álvaro	36	M	17/06/2024	6	NÃO	SIM
Rafael	20	M	10/07/2024	5	SIM	NÃO
Marina	25	F	03/09/2024	9	NÃO	NÃO
Fábio	**45**	**M**	**10/08/2024**	**3**	**NÃO**	**???**

Exemplo - Classificação

Coleta de Dados

Dados de Entrada: Informações sobre os clientes, como idade, gênero, histórico de compras, frequência de uso do serviço, interações com o suporte ao cliente, etc.

Dados de Saída: Um rótulo indicando se o cliente deixou o serviço (churn) ou não (não churn).

Pré-processamento

Limpeza de Dados: Remover valores ausentes ou inconsistentes.

Transformação de Dados: Converter variáveis categóricas em numéricas (por exemplo, gênero: masculino = 0, feminino = 1).

Normalização: Escalar os dados para que todas as variáveis tenham a mesma importância.

Divisão de Dados

Dividir o conjunto de dados em um conjunto de treinamento (80%) e um conjunto de teste (20%).

Exemplo - Classificação

Treinamento

Escolher um algoritmo de classificação, como a Árvore de Decisão, Random Forest ou Logistic Regression. Treinar o modelo usando o conjunto de treinamento. O modelo aprenderá a associar as características dos clientes com a probabilidade de churn.

Avaliação

Avaliar o desempenho do modelo usando o conjunto de teste. Métricas comuns incluem Acurácia, Precisão, Recall e F1-Score.

Ajuste

Ajustar hiperparâmetros para melhorar o desempenho do modelo. Utilizar técnicas como Validação Cruzada para garantir que o modelo não esteja superajustado aos dados de treinamento.

Predição

Usar o modelo treinado para prever a probabilidade de churn para novos clientes.

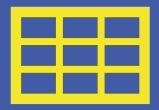
Exemplo - Regressão

Imagine que você tenha um conjunto de dados de casas, onde cada casa tem características como tamanho, número de quartos, localização, etc., e um preço associado. O objetivo seria prever o preço de uma casa...

Variável target

ID casa	Tamanho (m²)	Quantidade de quartos	UF de localização	Bairro de localização	Distância para o aeroporto (km)	Preço (R$)
001	360	3	GO	Valparaíso de Goiás	40	300.000
002	120	1	DF	Samambaia	30	250.000
003	250	2	CE	Paracuru	300	220.000
004	45	2	DF	Noroeste	20	500.000
005	**100**	**2**	**DF**	**Asa Sul**	**10**	**???**

Exemplo - Regressão



Coleta de Dados

1000 casas com suas características (dados de entrada) e preços (dados de saída).

Pré-processamento

Os dados precisam ser limpos e transformados. Isso pode incluir a remoção de valores ausentes, transformação de variáveis categóricas em numéricas e normalização de dados (por exemplo, transformar metros quadrados em uma escala comum).

Divisão de Dados

O conjunto de dados é dividido em duas partes: um conjunto de treinamento e um conjunto de teste. O conjunto de treinamento é usado para treinar o modelo, enquanto o conjunto de teste é usado para avaliar seu desempenho. Exemplo: 800 casas para treinamento, 200 para teste.

Exemplo - Regressão

Treinamento

O modelo é treinado usando o conjunto de treinamento. Durante o treinamento, o modelo aprende a associar as características dos dados de entrada (por exemplo, tamanho da casa, número de quartos) com os valores de saída (preço da casa). Como exemplo podemos usar um algoritmo como Regressão Linear para aprender a prever o preço com base nas características.

Avaliação

Após o treinamento, o modelo é avaliado usando o conjunto de teste. Métricas como erro quadrático médio (MSE), erro absoluto médio (MAE) e R^2 são usadas para medir o desempenho do modelo.

Ajuste

Com base na avaliação, o modelo pode ser ajustado para melhorar seu desempenho (reduzir o erro). Isso pode incluir a alteração de hiperparâmetros ou a utilização de técnicas de regularização.

Predição

Finalmente, o modelo treinado pode ser usado para prever valores contínuos para novos dados não rotulados (Prever o preço de novas casas com base em suas características).

Exemplos de algoritmos

Classificação

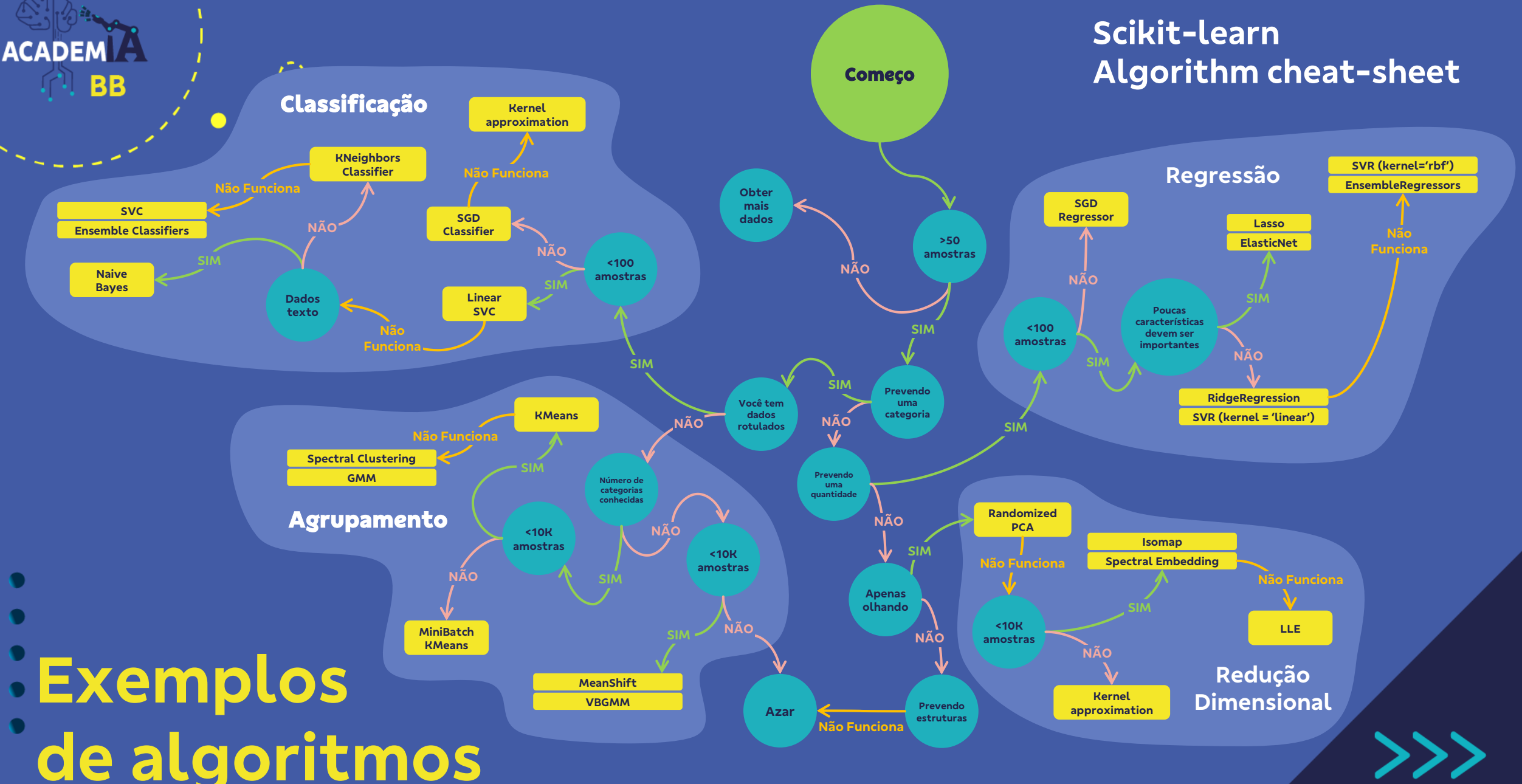
- Gradient Boosting
- Support Vector Machines (SVM)
- Logistic Regression
- Decision Tree
- Random Forest
- KNN

Regressão

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting

ACADEMIA BB

Scikit-learn Algorithm cheat-sheet



Exemplos de algoritmos

Exemplos de Métricas - Classificação

Matriz de Confusão

Mostra a contagem de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos.

		Valor Predito	
		Não	Sim
Real	Não	Verdadeiro Negativo (TN)	Falso Positivo (FP)
	Sim	Falso Negativo (FN)	Verdadeiro Positivo (TP)

Exemplos de Métricas - Classificação

Acurácia

Proporção de previsões corretas sobre o total de previsões.

Precisão

Proporção de verdadeiros positivos sobre o total de positivos preditos.

Recall

Proporção de verdadeiros positivos sobre o total de positivos reais.

F1 Score:

Média harmônica da precisão e do recall, útil quando há um desbalanceamento entre as classes.

Curva ROC e AUC

A curva ROC mostra a taxa de verdadeiros positivos contra a taxa de falsos positivos. A AUC é a área sob a curva ROC e fornece uma medida agregada de desempenho em todas as possíveis classificações.

Exemplos de Métricas - Regressão

Erro Médio Absoluto (MAE - Mean Absolute Error)

Descrição: Mede a média das diferenças absolutas entre os valores previstos e os valores reais.

Interpretação: Um MAE menor indica um modelo mais preciso.

Erro Quadrático Médio (MSE - Mean Squared Error)

Descrição: Mede a média dos quadrados das diferenças entre os valores previstos e os valores reais.

Interpretação: Penaliza erros maiores mais severamente do que o MAE.

Raiz do Erro Quadrático Médio (RMSE - Root Mean Squared Error)

Descrição: É a raiz quadrada do MSE, trazendo a métrica de volta à mesma unidade dos valores previstos.

Interpretação: Facilita a interpretação dos erros em termos das unidades originais dos dados.

R-quadrado (R^2 - Coeficiente de Determinação)

Descrição: Mede a proporção da variância na variável dependente que é previsível a partir das variáveis independentes.

Interpretação: Um valor de R^2 próximo de 1 indica um bom ajuste do modelo aos dados.

R-quadrado Ajustado (Adjusted R^2)

Descrição: Ajusta o R^2 para o número de preditores no modelo, penalizando a adição de variáveis irrelevantes.

Interpretação: Útil para comparar modelos com diferentes números de preditores

Relembrando...

Nas primeiras lives aprendemos a fazer a coleta de dados e o pré-processamento. Alguns comandos que aprendemos:

Carregar dados:

read_csv, head, info, value_counts, columns, copy

Exploratória e transformação de dados:

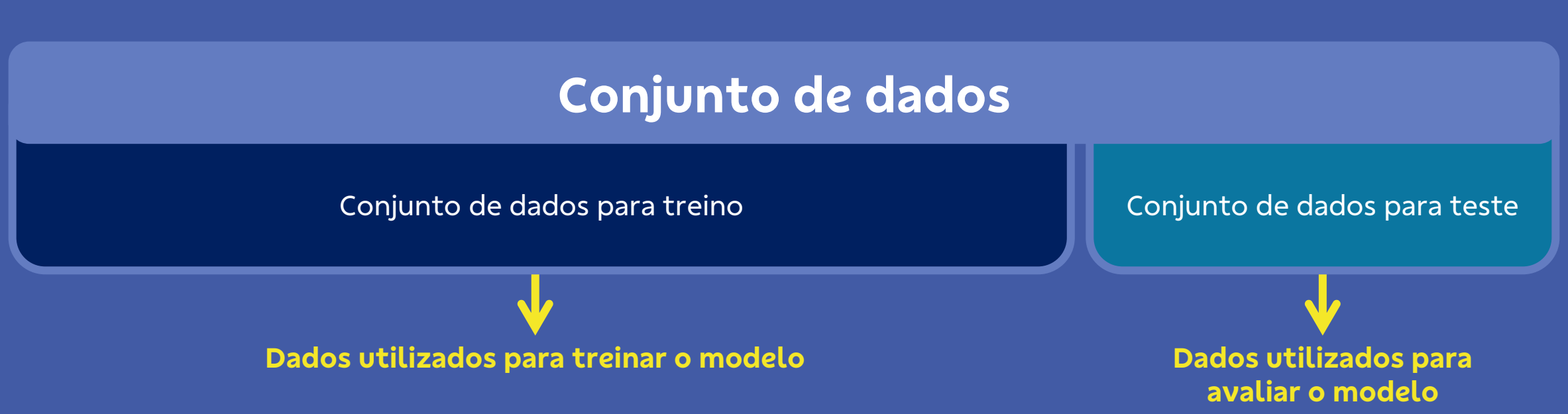
to_datetime, isna, fillna, corr, describe, get_dummies

Normalização:

scaler

Então agora vamos partir da divisão em treino e teste, ok?

Divisão treino/teste – holdout



Divisão treino/teste – holdout

Dados de Treino

Usados para treinar o modelo, ou seja, para ajustar os parâmetros do algoritmo.

Geralmente, cerca de 70-80% dos dados totais são usados para treino.

O modelo aprende a partir desses dados, ajustando-se para minimizar o erro nas previsões.

Dados de Teste

Usados para avaliar a performance do modelo em dados que ele não viu durante o treino.

Normalmente, cerca de 20-30% dos dados totais são reservados para teste.

Após o treinamento, o modelo é testado com esses dados para verificar sua capacidade de generalização e identificar possíveis problemas como overfitting (quando o modelo se ajusta demais aos dados de treino e não generaliza bem para novos dados).

Importância da Divisão

Generalização: A divisão garante que o modelo não apenas memorize os dados de treino, mas também consiga fazer previsões precisas em novos dados.

Avaliação Justa: Permite uma avaliação justa da performance do modelo, já que os dados de teste não foram usados durante o treinamento.

Mão na Massa

ACADEMIA BB

Cronograma IA e Ciência de Dados

> Lives de Conteúdo

1ª Live: Desvendando o CRISP-DM: Tipos de Dados e Variáveis para Iniciantes - **03/10 às 9h**
2ª Live: Exploração de Dados e Engenharia de Variáveis: O Segredo dos Cientistas de Dados - **08/10 às 9h**
3ª Live: Aprendizado de Máquina Não-Supervisionado: Descubra Padrões Ocultos - **10/10 às 9h**
4ª Live: Aprendizado de Máquina Supervisionado: Encontrando respostas a partir de padrões- **15/10 às 9h**

> Mentorias

17/10 - 22/10 - 24/10 - 29/10
Todas as sessões ocorrerão **às 9h e às 14h**

> Desafio

15/10: Início do prazo para desenvolvimento do desafio
15/11: Fim do prazo para entrega do desafio

Outubro

D	S	T	Q	Q	S	S
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

Novembro

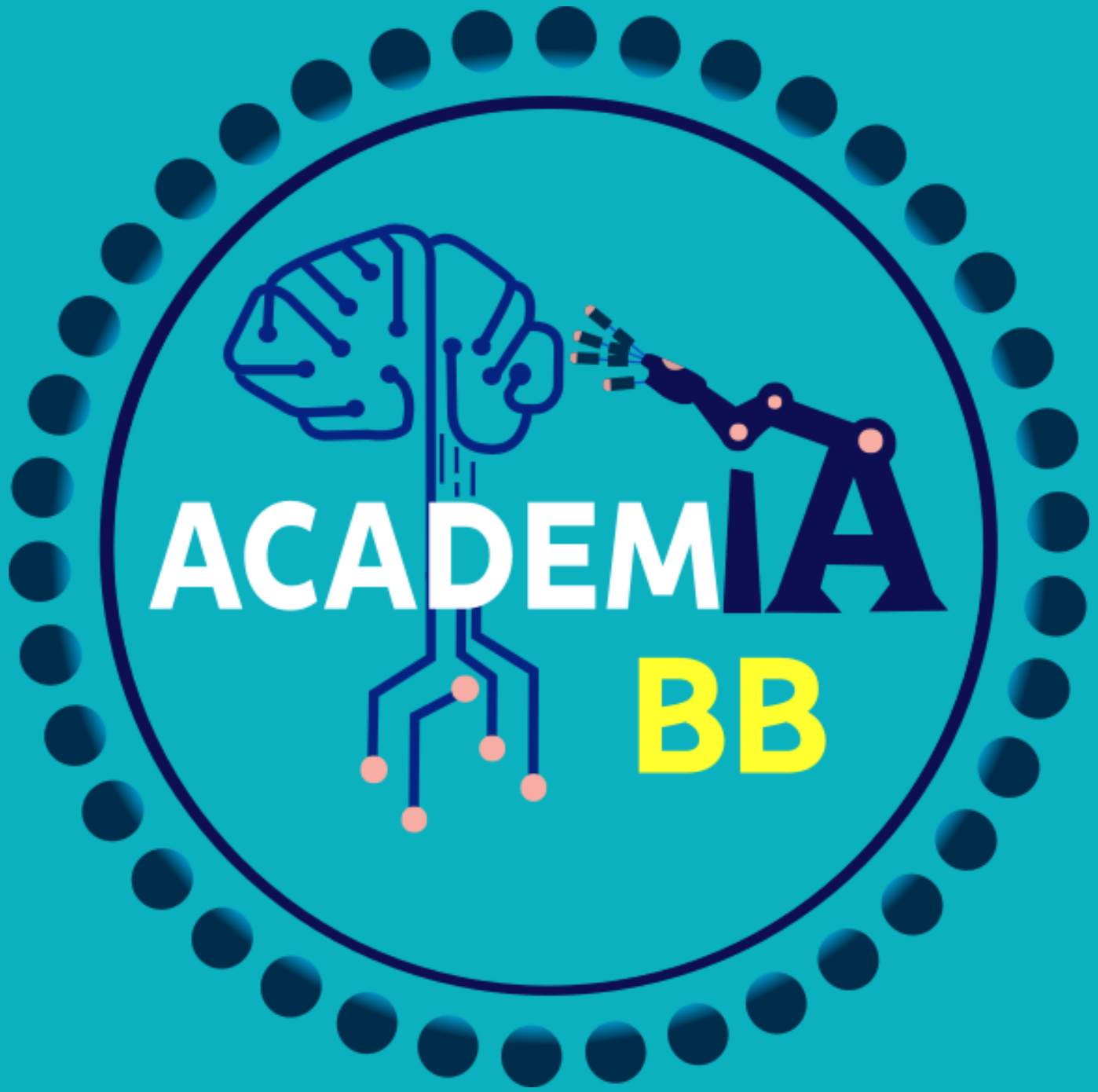
D	S	T	Q	Q	S	S
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30/31

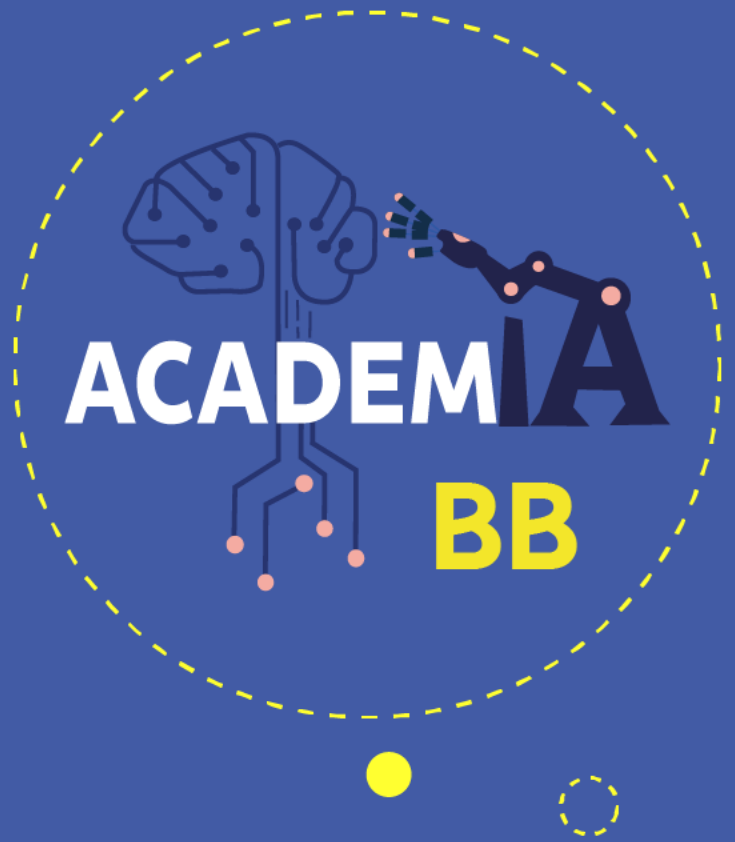
Lives de conteúdo | Mentorias | Início e término do prazo para o desafio

UniBB

Obrigado(a)!

BB DATA DRIVEN UniBB


ACADEMIA
BB



Bibliografia

https://www.ibm.com/br-pt/topics/supervised-learning;

Curso Alura 201799 - Árvores de Decisão: aprofundando em modelos de Machine Learning;

https://cursa.app/pt/pagina/principios-de-aprendizado-supervisionado-conjuntos-de-dados-treino-e-teste;

https://scikit-learn.org/stable/index.html;