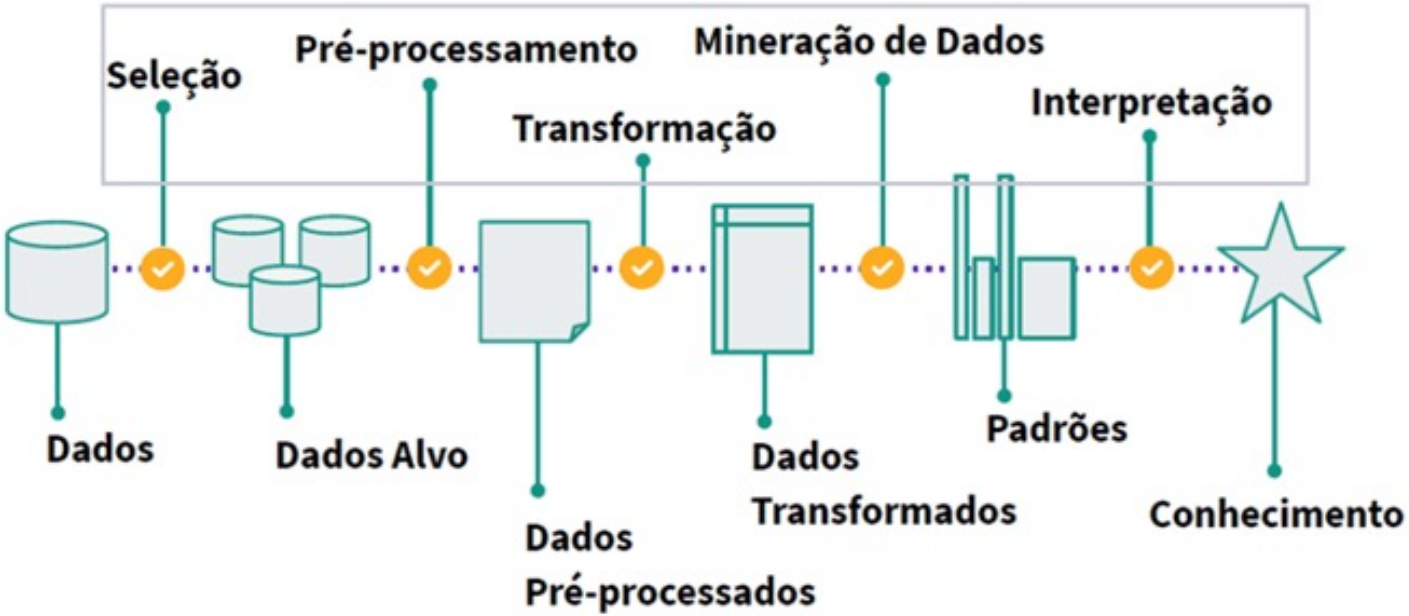
MINERAÇÃO DE DADOS

Prof. Me. Napoleão Póvoa Ribeiro Filho

UNITINS

GOVERNO DO TOCANTINS
TRABALHANDO E CUIDANDO DE TODOS

MINERAÇÃO DE DADOS



APRENDIZADO DE MÁQUINA

- Supervisionado
  - Regressão
  - Classificação
- Não-supervisionado
  - Agrupamento (clustering)
  - Redução de dimensionalidade
- Aprendizado por reforço
  - Agente
  - Ambiente
  - Política

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

ALGORITMOS DE APRENDIZADO DE MÁQUINA

Aprendizado Supervisionado

- **Regression**
  - Linear Regression
- **Classification**
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - k-Nearest Neighbors (k-NN)
  - Support Vector Machine (SVM)
  - Naive Bayes

ALGORITMOS DE APRENDIZADO DE MÁQUINA

Aprendizado Não-Supervisionado

- **Clustering**
  - k-Means
  - DBSCAN
  - Hierarchical Clustering
- **Dimensionality Reduction**
  - Principal Component Analysis (PCA)
  - t-SNE
  - UMAP

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

ALGORITMOS DE APRENDIZADO DE MÁQUINA

Aprendizado por Reforço

- Q-Learning
- SARSA
- Deep Q-Networks (DQN)
- Policy Gradient Methods (ex: REINFORCE, PPO)

CLASSIFICAÇÃO

- A classificação é uma forma de análise de dados que extrai modelos para descrever importantes classes de dados
- Esses modelos, chamados classificadores, predizem rótulos de classe categóricos (discretos, não ordenados)
- Exemplo: Categorizar aplicações de empréstimo bancário como seguras ou arriscadas

CLASSIFICAÇÃO

- A classificação prediz rótulos de classe categóricos, enquanto a predição numérica (regressão) prediz valores contínuos ou ordenados.
- Possui inúmeras aplicações, incluindo detecção de fraude, marketing direcionado, previsão de desempenho, manufatura e diagnóstico médico

CONCEITOS BÁSICOS

- **Rótulo de Classe**: Atributo com valores discretos e não ordenados que representam as categorias ou classes. Exemplo: "seguro" ou "arriscado", "sim" ou "não", "tratamento A", "tratamento B" ou "tratamento C“
- **Atributos**: Características usadas para descrever os dados. Um tupla (ou exemplo) é representada por um vetor de atributos

CONCEITOS BÁSICOS

- **Dados de Treinamento:** Tuplas de dados com seus rótulos de classe associados, usados para construir o modelo de classificação
- **Dados de Teste**: Tuplas de dados independentes dos dados de treinamento, usados para estimar a precisão do modelo

PROCESSO DE CLASSIFICAÇÃO

- Etapa 1: **Construção do modelo**
  - Um algoritmo de classificação analisa o conjunto de dados de treinamento
  - Constrói um classificador (modelo) que descreve um conjunto predeterminado de classes.
  - O modelo pode ser representado na forma de regras de classificação, árvores de decisão ou fórmulas matemáticas

PROCESSO DE CLASSIFICAÇÃO

- Etapa 1: **uso do modelo (classificação)**
  - A precisão preditiva do classificador é estimada usando um conjunto de dados de teste
  - Se a precisão for aceitável, o modelo pode ser usado para classificar novos dados para os quais o rótulo de classe é desconhecido

EXEMPLOS

TAREFA	CONJUNTO DE ATRIBUTOS (x)	RÓTULO DA CLASSE (y)
Classificação de e-mails como spam	Frequência de palavras, presença de links, remetente, horário de envio	Spam ou Não Spam
Diagnóstico de diabetes	Glicose, pressão arterial, índice de massa corporal, idade	Diabético ou Não Diabético
Classificação de sentimentos em avaliações	Palavras do texto, número de caracteres, presença de emojis	Positivo, Neutro ou Negativo

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

TÉCNICAS DE CLASSIFICAÇÃO

- Classificadores base (simples e fáceis de interpretar)
  - Árvore de decisão
  - Métodos baseados em regras
  - Nearest-neighbor (Vizinho-mais-próximo)
  - NaïveBayes
  - Redes neurais
  - SupportVector Machines (SVM)
  - Deep Learning
- Classificadores de assembleia (combinação de vários classificadores)
  - Boosting, Bagging, Random Forests

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

EXEMPLO ÁRVORE DE DECISÃO

Idade	Renda mensal	Score de crédito	Tempo de emprego	Valor do empréstimo	Possui dívidas?	Aprovado?
35	5000	750	5	10000	não	sim
22	2000	520	1	8000	sim	não
40	9000	820	10	20000	não	sim
29	3500	610	2	15000	sim	não
50	6000	700	15	12000	não	sim

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

ÁRVORE DE DECISÃO

- Tem esse nome porque começa com um nó raiz (root) e se ramifica em nós filhos, formando um caminho de decisão até o resultado final.
- É um modelo de classificação (ou regressão) que organiza decisões em uma estrutura hierárquica parecida com um fluxograma
- Cada nó interno representa uma condição sobre um atributo, cada ramo representa um resultado da condição, e cada nó folha representa uma classe final (rótulo)

TRANSFORMAÇÃO

- Algoritmos como árvores de decisão não entendem texto: eles precisam de números para fazer cálculos de entropia, ganho de informação ou índice Gini.
  - Atributos categóricos são transformados em valores numéricos.
    - “não” = 0
    - “sim” = 1
  - Dados ausentes são tratados.

PRÉ-PROCESSAMENTO

Idade	Renda mensal	Score de crédito	Tempo de emprego	Valor do empréstimo	Possui dívidas?	Aprovado?
35	5000	750	5	10000	0	1
22	2000	520	1	8000	1	0
40	9000	820	10	20000	0	1
29	3500	610	2	15000	1	0
50	6000	700	15	12000	0	1

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

ÁRVORE DE DECISÃO - ETAPAS

- Seleção dos dados
- Pré-processamento
- Escolha do atributo raiz (critério de divisão)
- Divisão dos dados
- Recursividade
- Podas (opcional)
- Uso do modelo

ESCOLHA DO ATRIBUTO RAIZ

- Determina qual atributo usar primeiro (como raiz) para começar a separar os dados.
- Ela faz isso avaliando qual atributo separa melhor as classes (ex: “sim” ou “não”).
- Para isso, usamos medidas de seleção de atributos que indicam "quão puro" ou "informativo" um atributo é.

ESCOLHA DO ATRIBUTO RAIZ

- **Ganho de informação** (Information Gain)
- **Entropia**: é uma medida de "bagunça": quanto mais misturado estiver (sim/não), maior a entropia.
- Quanto maior o ganho de informação, melhor o atributo separa as classes (menor o valor da entropia).
  - Entropia = 1: máxima incerteza (ex: 50% sim, 50% não)
  - Entropia entre 0 e 1: há mistura de classes (quanto mais perto de 0, mais puro)
  - Entropia = 0: 100% certeza (ótimo para árvores de decisão)

$$\text{Entropia} = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$$

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

ESCOLHA DO ATRIBUTO RAIZ

- Razão de Ganho (Gain Ratio)
- É uma versão ajustada do ganho de informação, usada no algoritmo C4.5
- Evita que o modelo escolha atributos com muitos valores distintos (como CPF ou ID), que tendem a dar ganho de informação alto, mas não generalizam bem.

ESCOLHA DO ATRIBUTO RAIZ

- **Índice de Gini (Gini Index)**
- Mede o grau de impureza de um conjunto de dados.
- Quanto menor o Gini, mais puro o grupo.
  - Se um grupo tem só “sim” ou só “não” → Gini = 0 (ótimo)
  - Se tem metade “sim” e metade “não” → Gini alto (ruim)
- Usado no algoritmo CART (Classification and Regression Tree).

ESCOLHA DO ATRIBUTO RAIZ

Medida	**O que faz?**	**Melhor valor possível**
Ganho de informação	Mede quanto a incerteza é reduzida	Quanto maior, melhor
Razão de Ganho	Ajusta o ganho para evitar escolhas tendenciosas	Quanto maior, melhor
Índice de Gini	Mede o quão misturados estão os dados	Quanto maior, melhor

RECURSIVIDADE

- O processo se repete em cada subconjunto, criando novos nós e ramos, até atingir um critério de parada, como:
  - Todos os exemplos em um nó são da mesma classe
  - A profundidade máxima foi atingida
  - Não há mais atributos úteis para dividir

PODA (PRUNING)

- Após a criação da árvore, ramos irrelevantes ou sensíveis a ruídos podem ser removidos (poda), tornando o modelo mais simples e generalizável.
- É feita para evitar que a árvore aprenda demais os detalhes dos dados de treinamento, incluindo ruídos ou exceções que não representam padrões reais (overfitting) .

QUANDO PODAR?

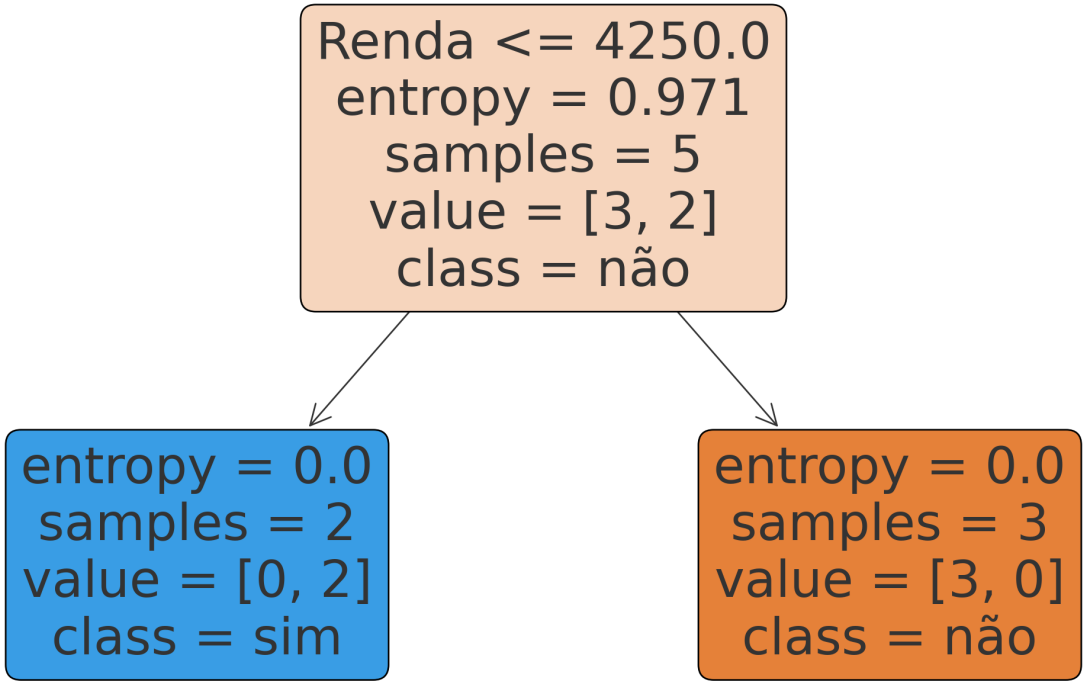
- Quando a árvore ficou muito complexa
  - Muitos ramos com poucos exemplos
  - A árvore está classificando muito bem os dados de treino, mas mal os dados novos
- Quando ramos não melhoram a precisão
  - Alguns ramos separam muito pouco ou quase nada
  - Mantê-los não melhora a performance do modelo
- Quando queremos um modelo mais simples e interpretável
  - Em contextos como medicina, educação ou negócios, simplicidade é fundamental

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

TIPOS DE PODA

- Pré-poda (pre-pruning)
  - Interrompe o crescimento durante a construção da árvore
  - Ex: "Só divida se o grupo tiver mais que 5 exemplos", ou "Pare quando o ganho de informação for pequeno“
- Pós-poda (post-pruning)
  - A árvore é construída até o fim, depois os ramos desnecessários são removidos
  - Ex: se um ramo representa um padrão fraco, ele é substituído por uma folha com a classe majoritária

ÁRVORE DE DECISÃO



UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS


UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS