BOOTCAMP

Python Aplicado à Análise de Dados e IA para o Setor Elétrico Brasileiro

IEEE | Universidade Federal do Rio Grande do Norte, IEEE Student Branch | IEEE PES Power & Energy Society

C2AI | CAECOMP | IEEE COMPUTER SOCIETY

Algoritmos Clássicos de Classificação

MIGUEL EURIPEDES NOGUEIRA DO AMARAL
miguel.amaral.101@ufrn.edu.br

QUEM SOU EU?

Miguel Euripedes

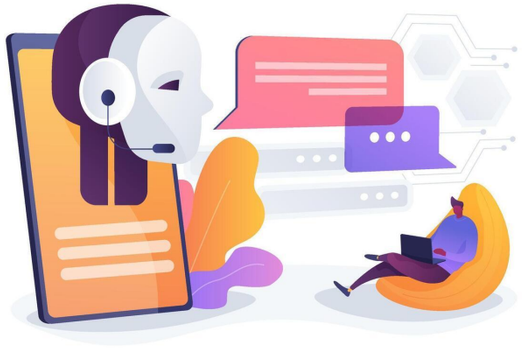
Bacharel em Ciências e Tecnologia - **UFRN**

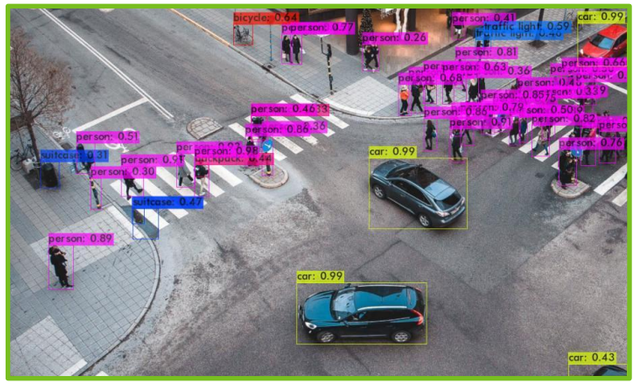
Engenheiro de Computação - **UFRN**

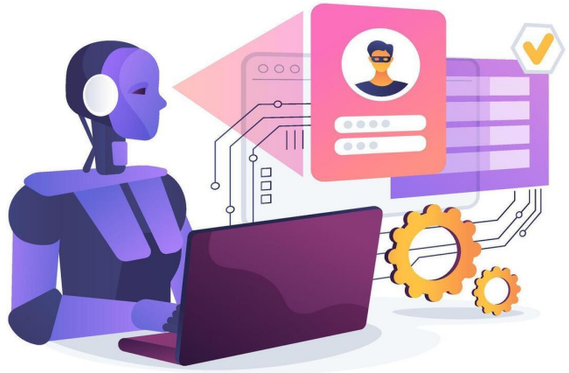
Mestrando Em Engenharia Elétrica e de Computação - **PPgEEC | UFRN**

Membro do grupo de pesquisa **Conect2ai** | Projetos: CNPq e Rota2030

IA?







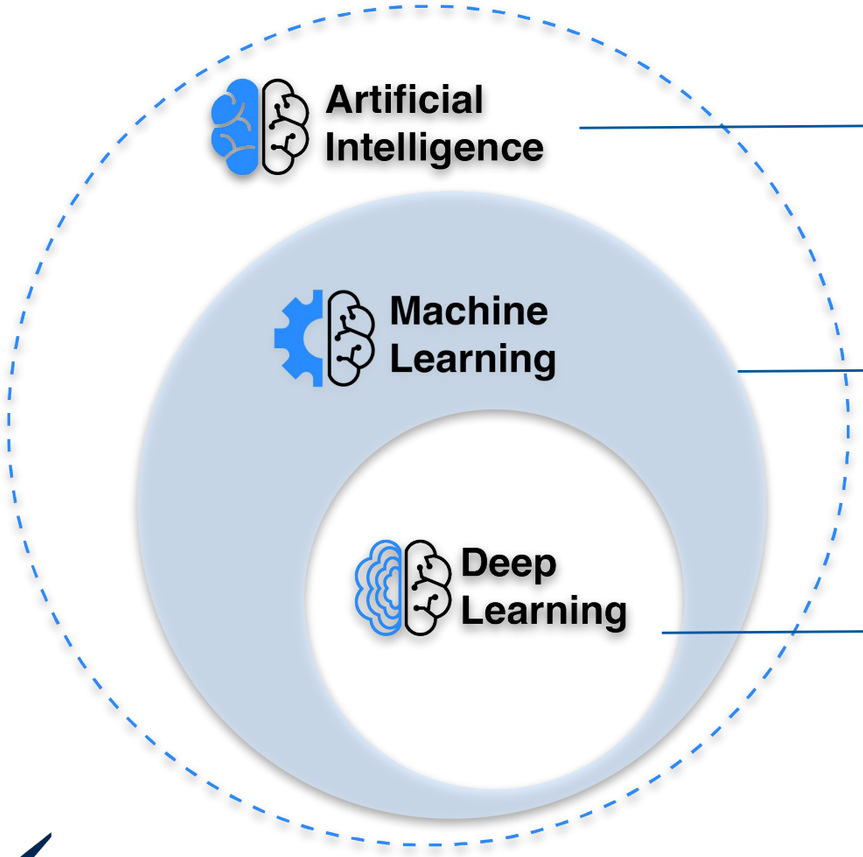
IEEE PES Power & Energy Society®

C2AI

CAECOMP

IEEE COMPUTER SOCIETY

PANORAMA GERAL DA IA



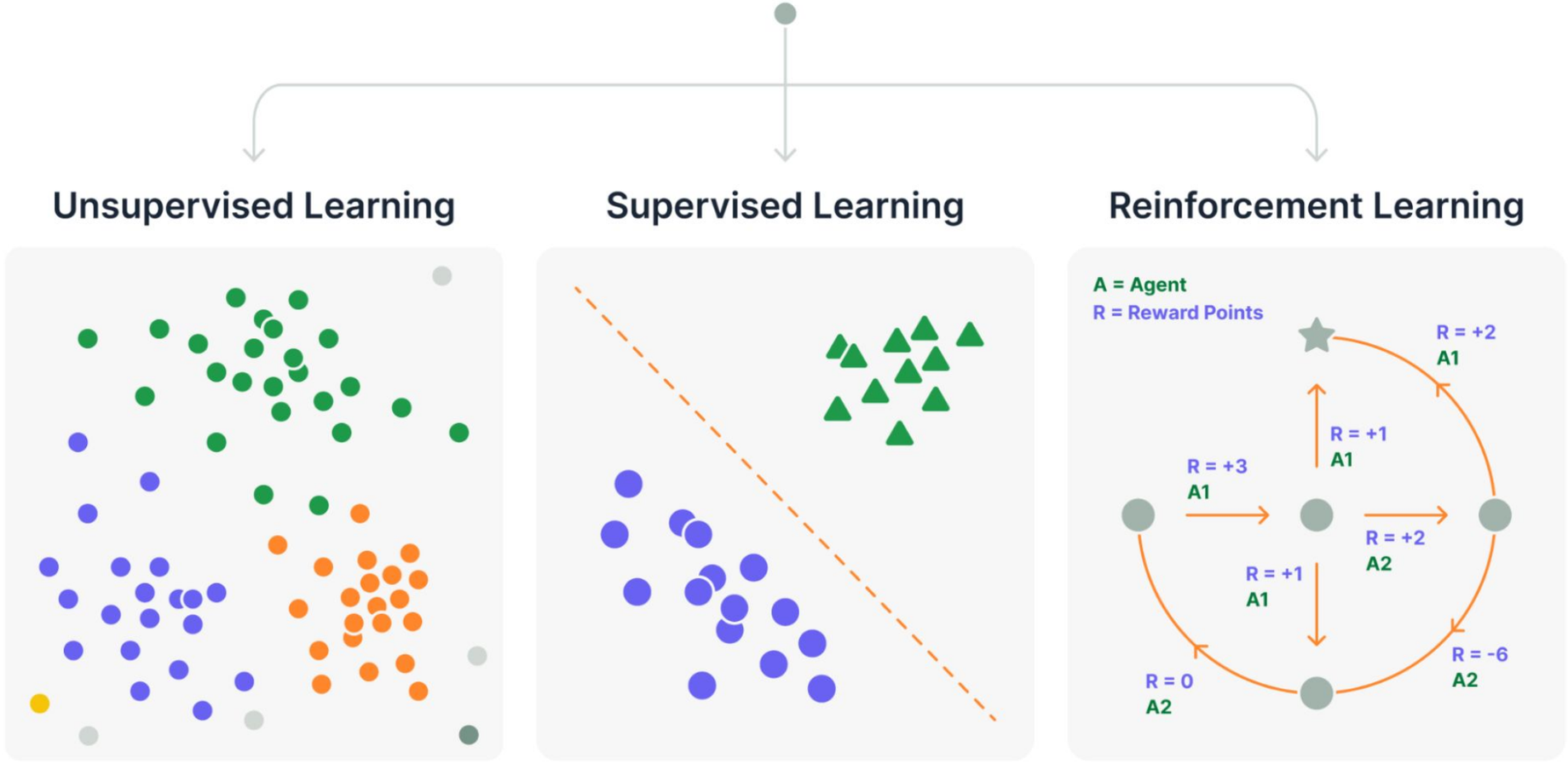
Área que busca criar sistemas capazes de simular comportamentos inteligentes.

Sistemas que aprendem padrões a partir de dados, em vez de depender apenas de programação explícita.

Técnicas baseadas em redes neurais profundas que permitem lidar com grandes volumes de dados e problemas complexos.

IEEE PES Power & Energy Society® | C2AI | CAECOMP | IEEE COMPUTER SOCIETY

PANORAMA GERAL DO APRENDIZADO DE MÁQUINA



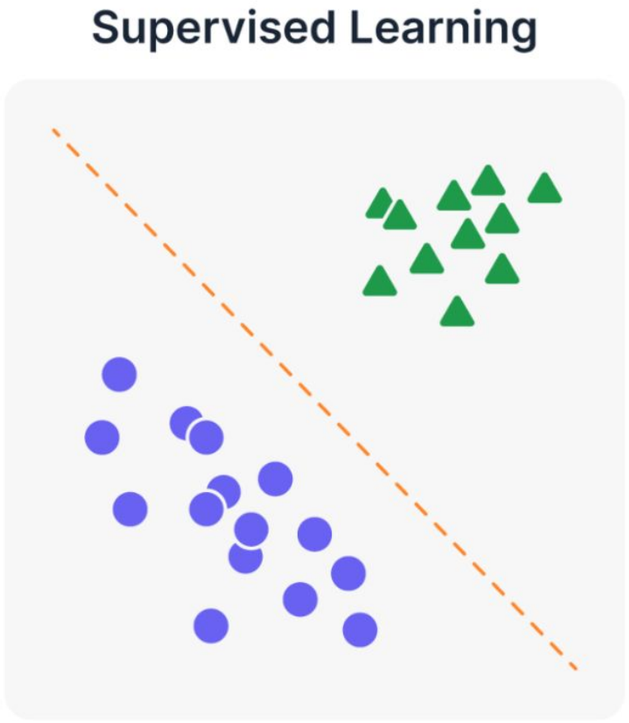
IEEE PES Power & Energy Society®

C2AI

CAECOMP

IEEE COMPUTER SOCIETY

APRENDIZADO SUPERVISIONADO



Aprendizado Supervisionado

Regressão	Classificação
Linear Regression	Logistic Regression
Ordinary Least Squares Regression	K-Nearest Neightbours
⋮	Decision Trees
	Random Forest
	Support Vector Machine
	⋮

IEEE PES Power & Energy Society®

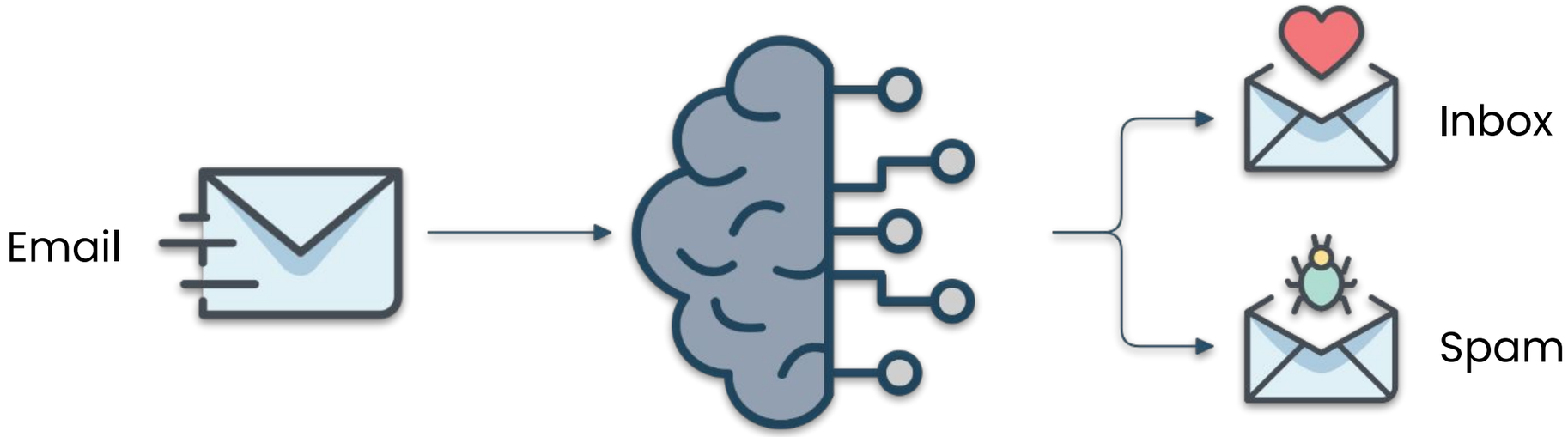
C2AI

CAECOMP

IEEE COMPUTER SOCIETY

O QUE É CLASSIFICAÇÃO?

A tarefa de atribuir um **rótulo** (uma categoria) a um dado



Modelo de Classificação

CONEXÃO COM A ENGENHARIA

Diagnóstico de Falhas → Sinais de sensores em máquinas ou torres de transmissão → Falha detectada ou Operação normal

- Manutenção de Aerogeradores → Vibração, ruído e velocidade de pás → Normal, Desgaste, Falha iminente

Controle de Processos → Parâmetros de produção (pressão, vazão, temperatura) → Estável, Instável, Crítico

- Qualidade da Energia → Tensão e frequência em parques eólicos/solares → Dentro da norma ou Instável

Gestão de Energia → Leituras de medidores inteligentes → Consumo Alto, Médio ou Baixo

IEEE PES Power & Energy Society®

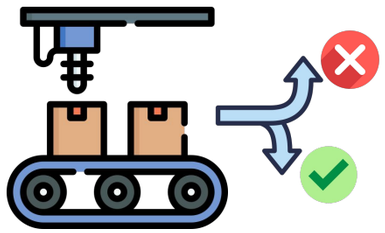
C2AI

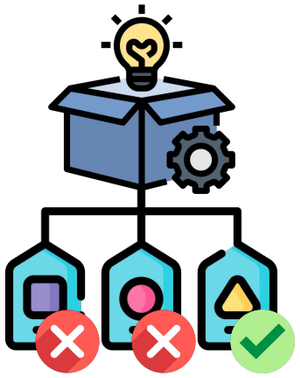
CAECOMP

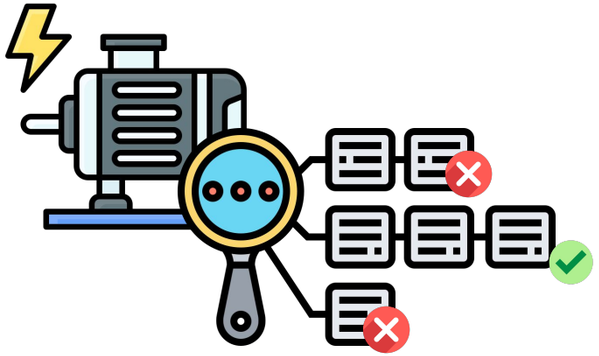
IEEE COMPUTER SOCIETY

TIPOS DE CLASSIFICAÇÃO

Binária	Multiclasse	Multirrótulo
Decide entre duas possibilidades	Escolhe uma única opção entre várias classes	Um item pode pertencer a mais de uma categoria ao mesmo tempo





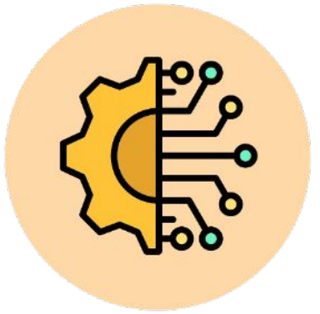


SEGREDO DO APRENDIZADO

Definir objetivo →

Selecionar o modelo →

Coleta de dados →



Treinar o Modelo →

Avaliar o Modelo

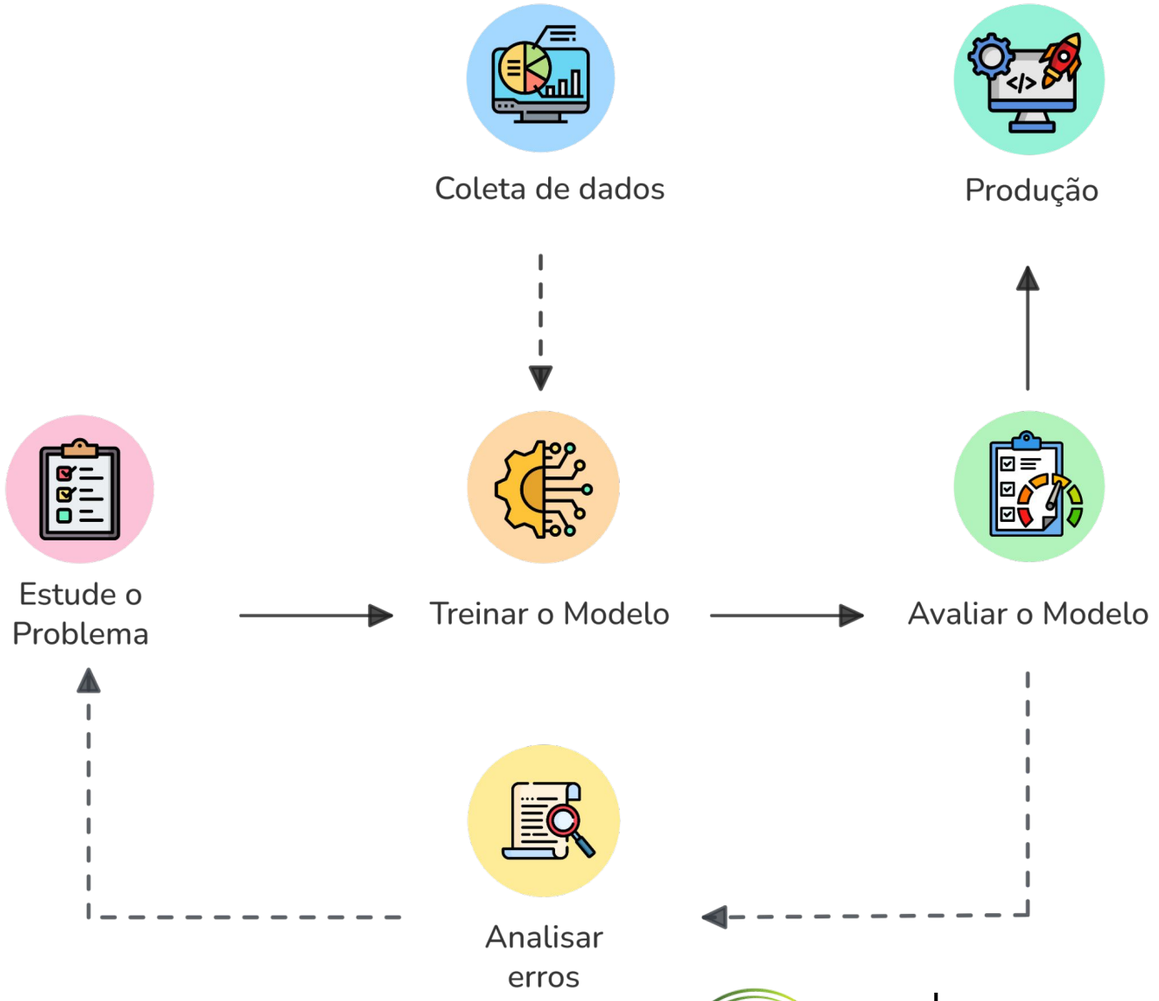
IEEE PES Power & Energy Society®

C2AI

CAECOMP

IEEE COMPUTER SOCIETY

SEGREDO DO APRENDIZADO



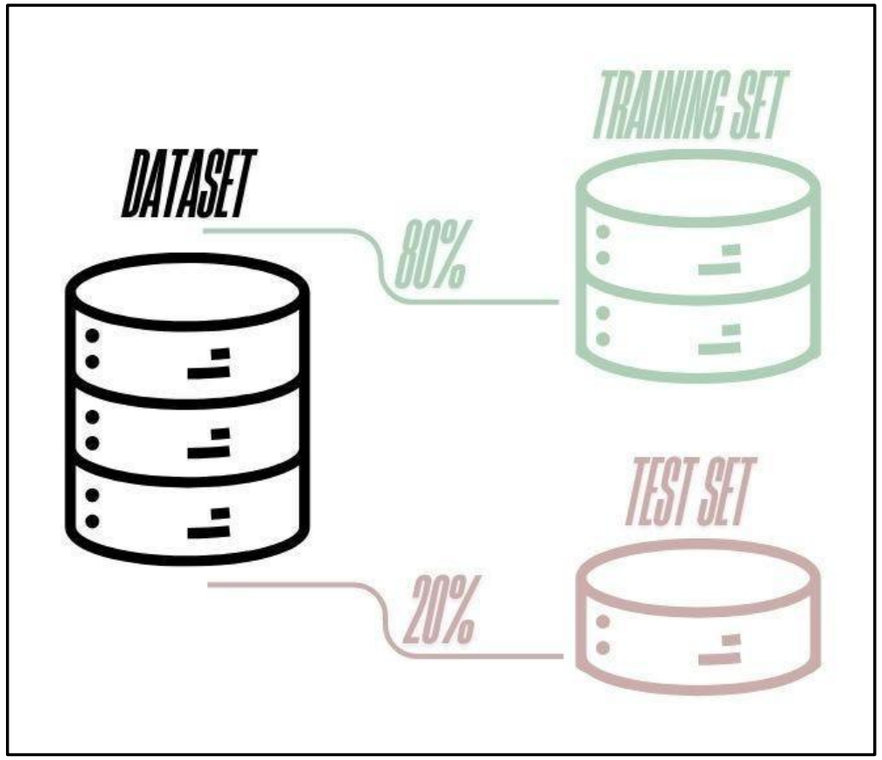
IEEE PES Power & Energy Society®

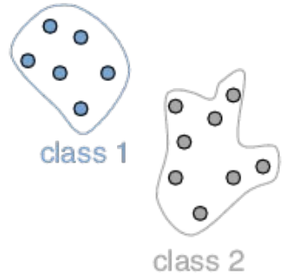
C2AI

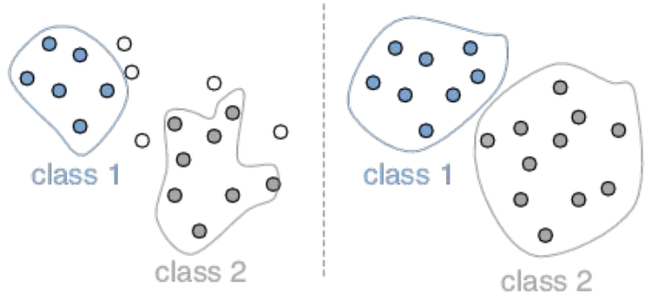
CAECOMP

IEEE COMPUTER SOCIETY

SEGREDO DO APRENDIZADO

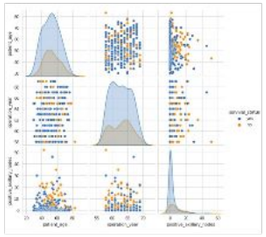




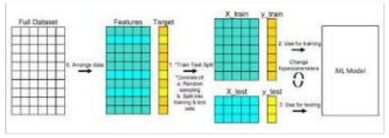


SEGREDO DO APRENDIZADO

Coleta e EDA dos Dados

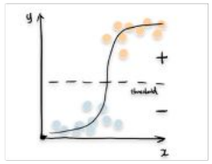


Pré-processamento e Separação Treino/Teste

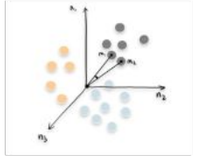


Treinamento

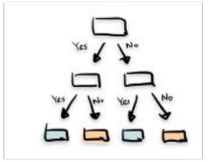
Regressão logística



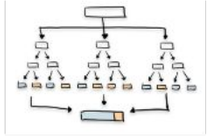
KNN



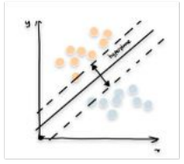
Árvore de Decisão



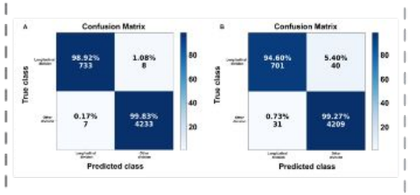
Floresta aleatória



SVM



Avaliação



IEEE PES Power & Energy Society®

C2AI

CAECOMP

IEEE COMPUTER SOCIETY

AVALIANDO MODELOS DE CLASSIFICAÇÃO

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = 0.70$$

$$\text{Recall} = \frac{TP}{TP + FN} = 0.80$$

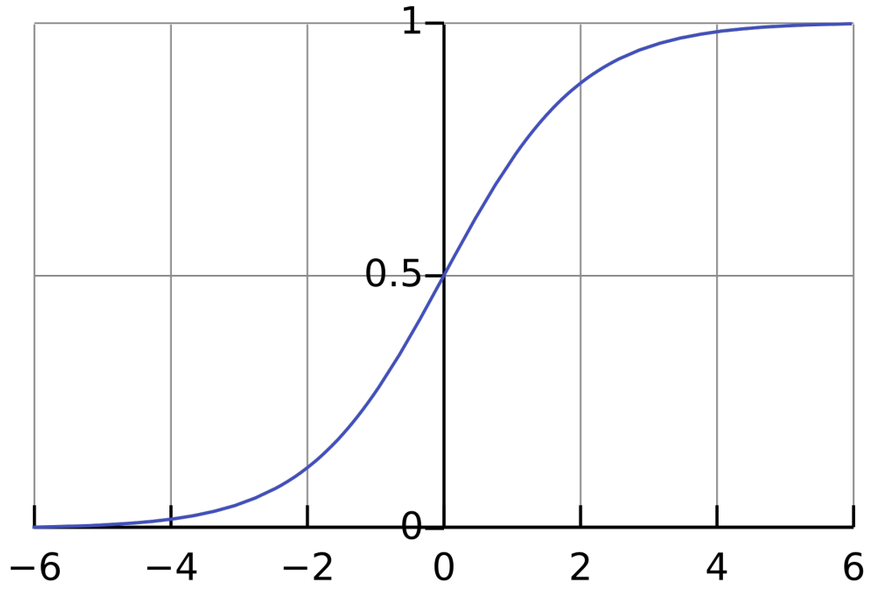
$$\text{Precision} = \frac{TP}{TP + FP} = 0.67$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = 0.73$$

MODELOS CLÁSSICOS

REGRESSÃO LOGÍSTICA

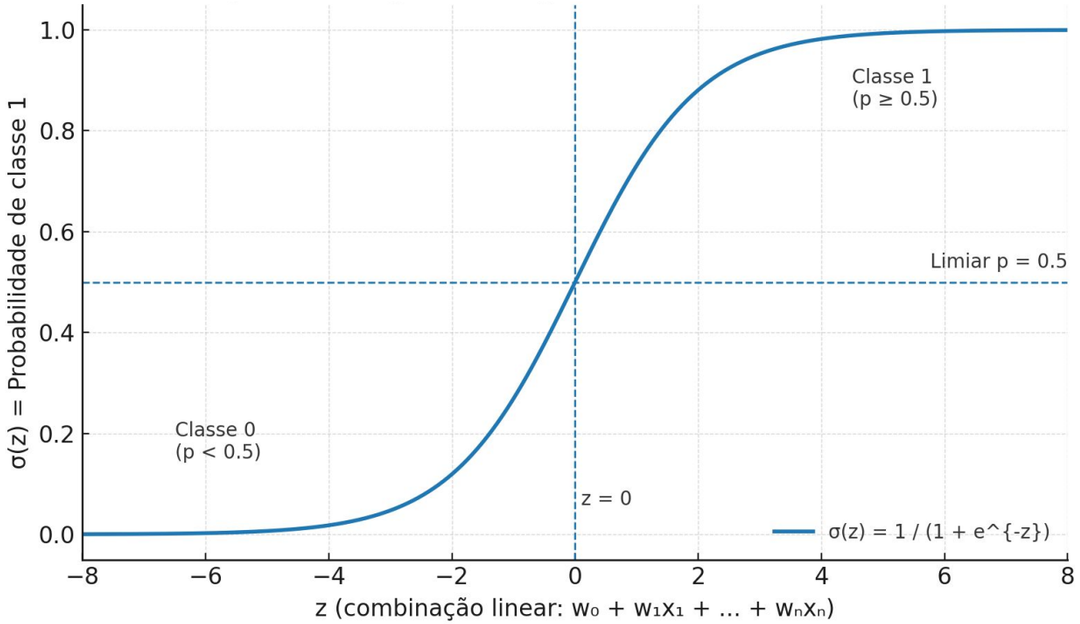
- Modelo de **classificação**, apesar do nome “regressão”→ **Classificação logística**?
- Objetivo: prever a **probabilidade de um evento** acontecer.
  - Para isso a saída é uma probabilidade entre **0 e 1**
- Para isso, usa a **função logística (sigmóide)**



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

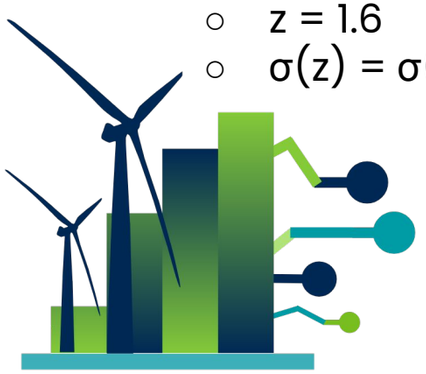
REGRESSÃO LOGÍSTICA

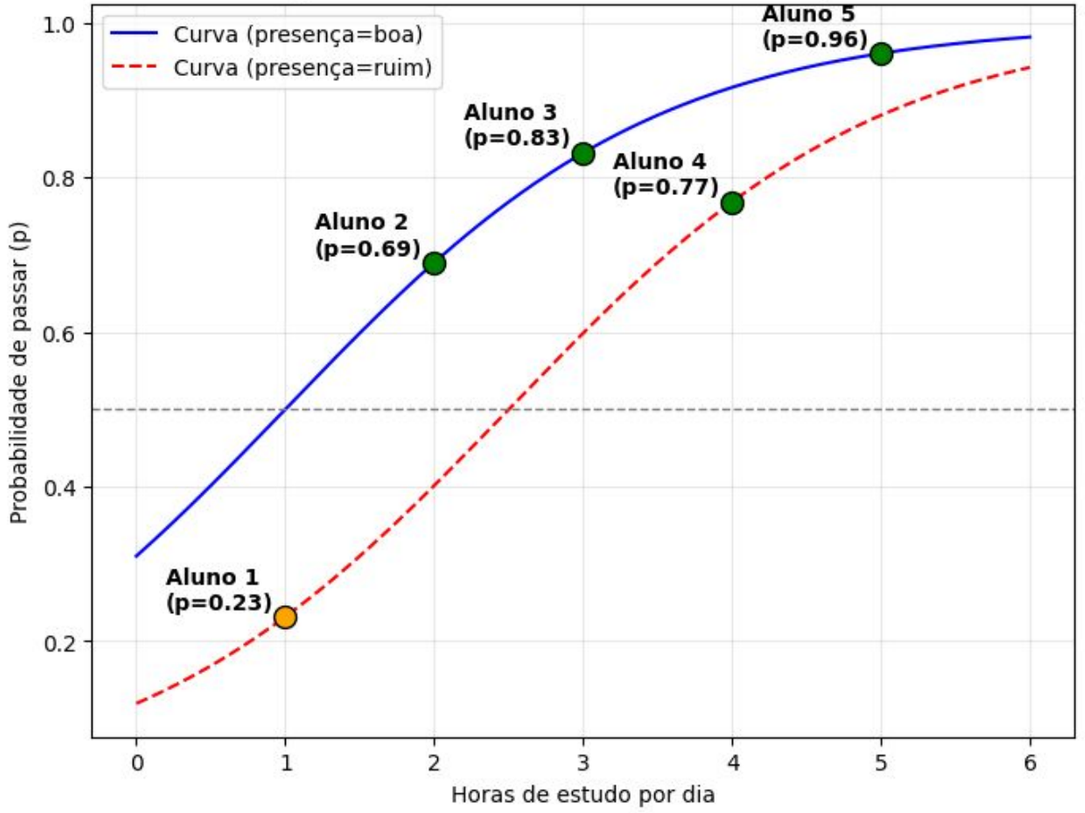
- $z = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n$ (combinação linear das variáveis).
- O resultado de $\sigma(z)$ é interpretado como **probabilidade de pertencer à classe 1**
- O modelo aprende os pesos (w) a partir dos dados, ajustando-os para **maximizar a chance de prever certo**.



REGRESSÃO LOGÍSTICA

- Dados:
  - Horas de estudo
  - Presença (1 = boa presença, 0 = ruim)
  - Média trabalhos, nota prova 1, etc.
  - **Rótulo:** 1 = passou e 0 = reprovado
- Modelo:
  - $z = w_0 + w_1(\text{horas}) + w_2(\text{presença}) + \ldots$
- **Suponha que, após treinar, o modelo aprendeu:**
  - $w_0 = -2$
  - $w_1(\text{horas}) = 0.8$
  - $w_2(\text{presença}) = 1.2$
- Para um aluno com **2 horas/dia** e **boa presença (1)**:
  - $z = -2{,}0 + 0{,}8 \times 3 + 1{,}2 \times 1$
  - $z = 1.6$
  - $\sigma(z) = \sigma(1.6) \approx 0.832$ (83.2%)





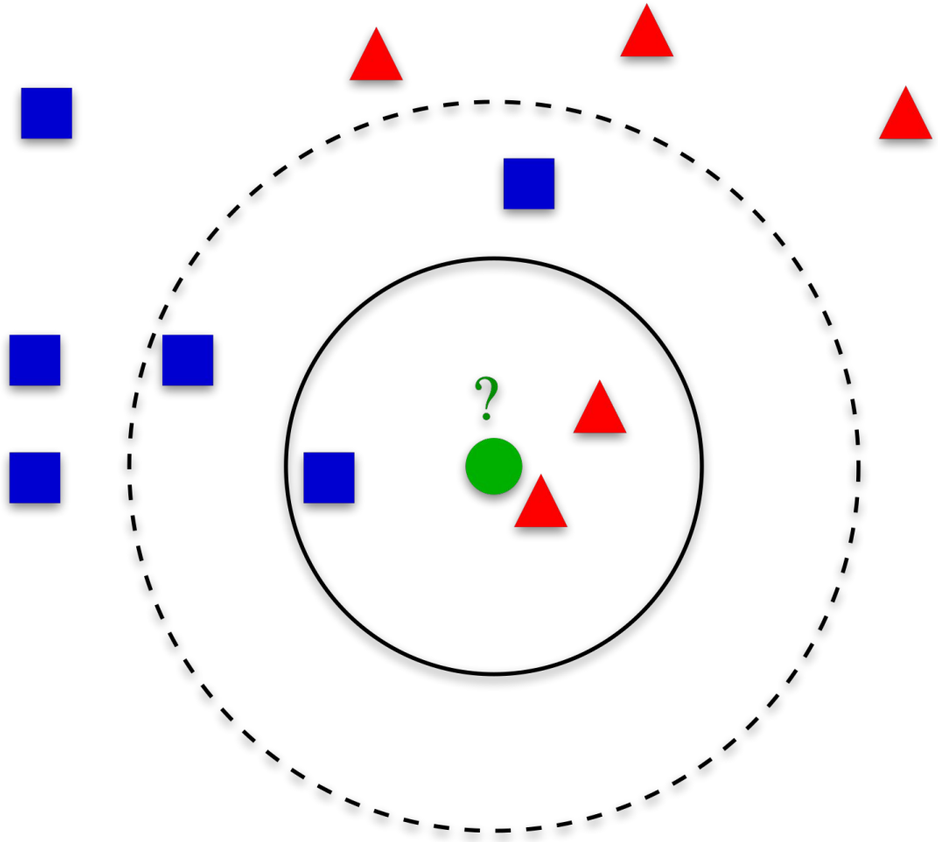
IEEE PES Power & Energy Society® | C2AI | CAECOMP | IEEE COMPUTER SOCIETY

KNN

- K-Nearest Neighbors → **K Vizinhos Mais Próximos**
- Não existe “equação” → o modelo **memoriza os dados de treino**.
- Objetivo: Classificar um novo exemplo com base na **semelhança** com exemplos anteriores

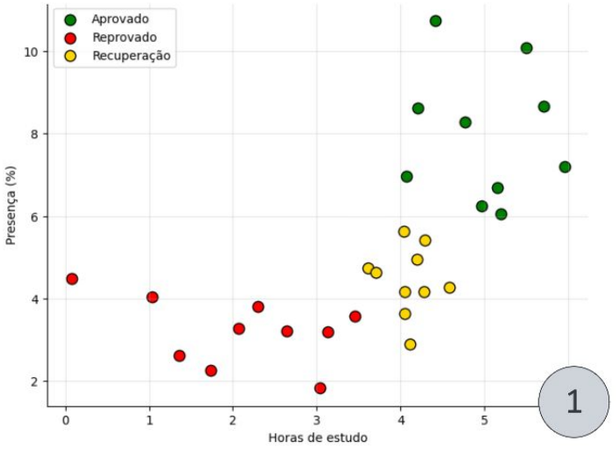
- Para classificar um novo aluno, o modelo:
  1. Calcula a **distância** entre o novo aluno e todos os alunos já conhecidos (ex.: distância Euclidiana).
  2. Pega os **K vizinhos mais próximos**.
  3. A classe mais frequente entre esses vizinhos é atribuída ao novo aluno.

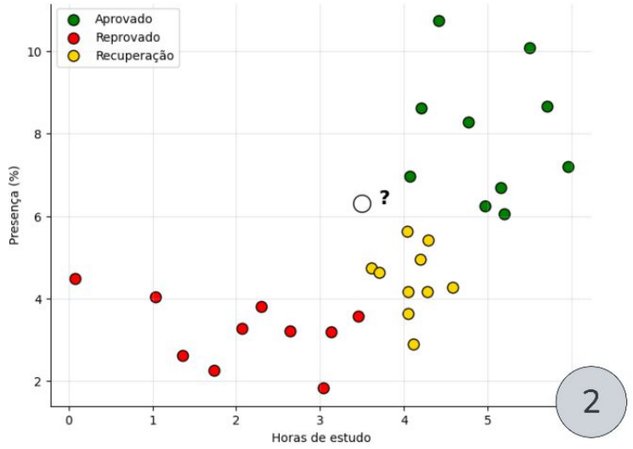
- Podemos usar 2 variáveis ou mais, o conceito de distância funciona no **espaço de n dimensões**.

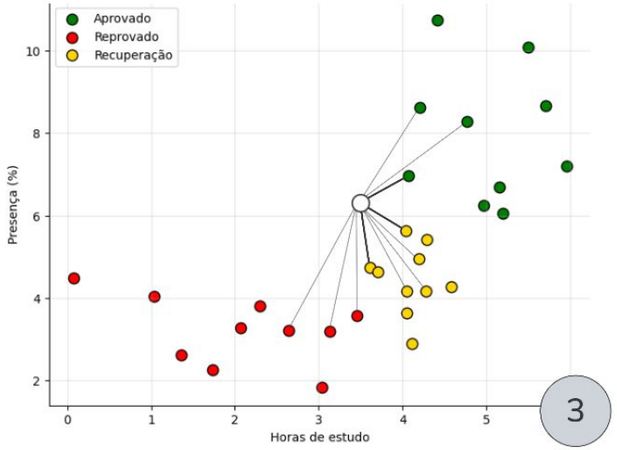


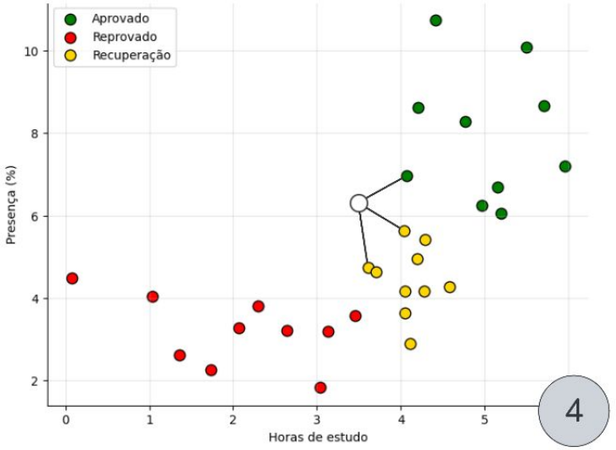
IEEE PES Power & Energy Society® | C2AI | CAECOMP | IEEE COMPUTER SOCIETY

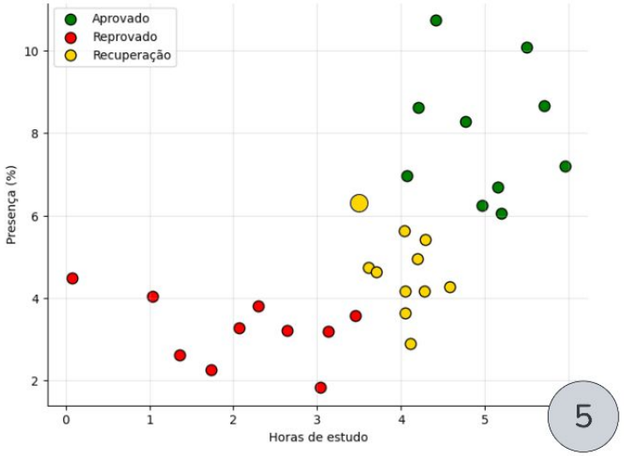
KNN









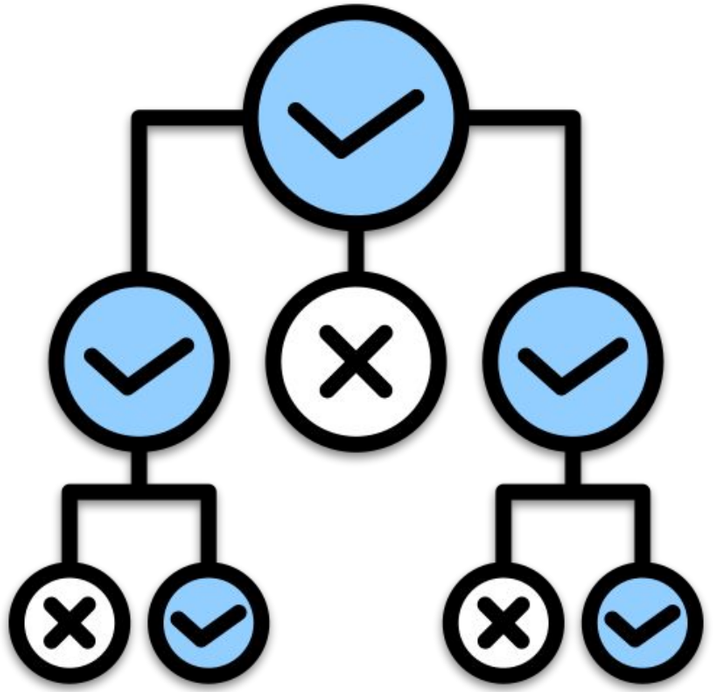


1 aprovados e 2 recuperação → novo aluno classificado como **recuperação**.

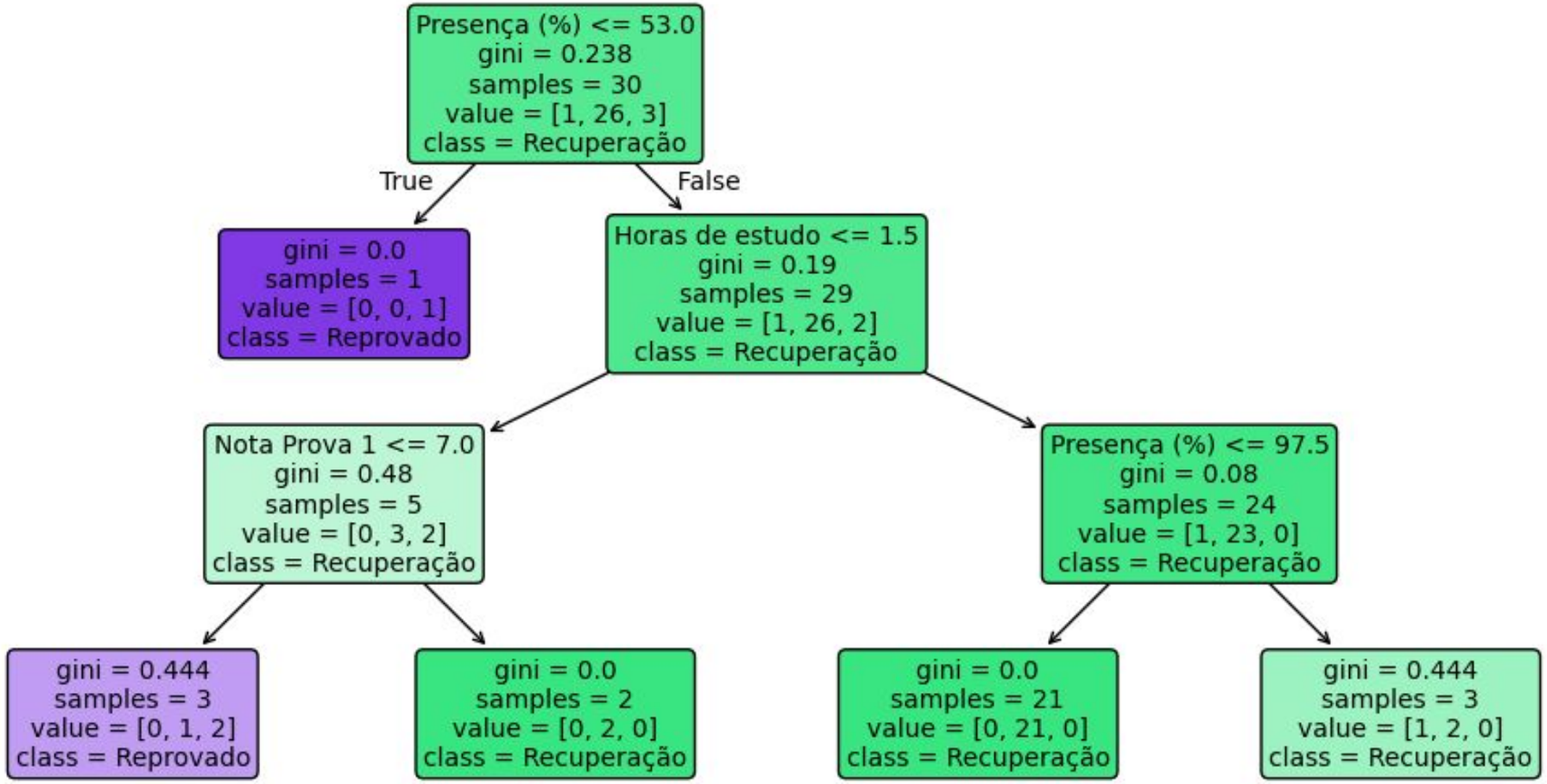
ÁRVORE DE DECISÃO

- **Objetivo:** Classificar um novo dado decidindo passo a passo com base em regras.
- A árvore é construída a partir dos **dados**.

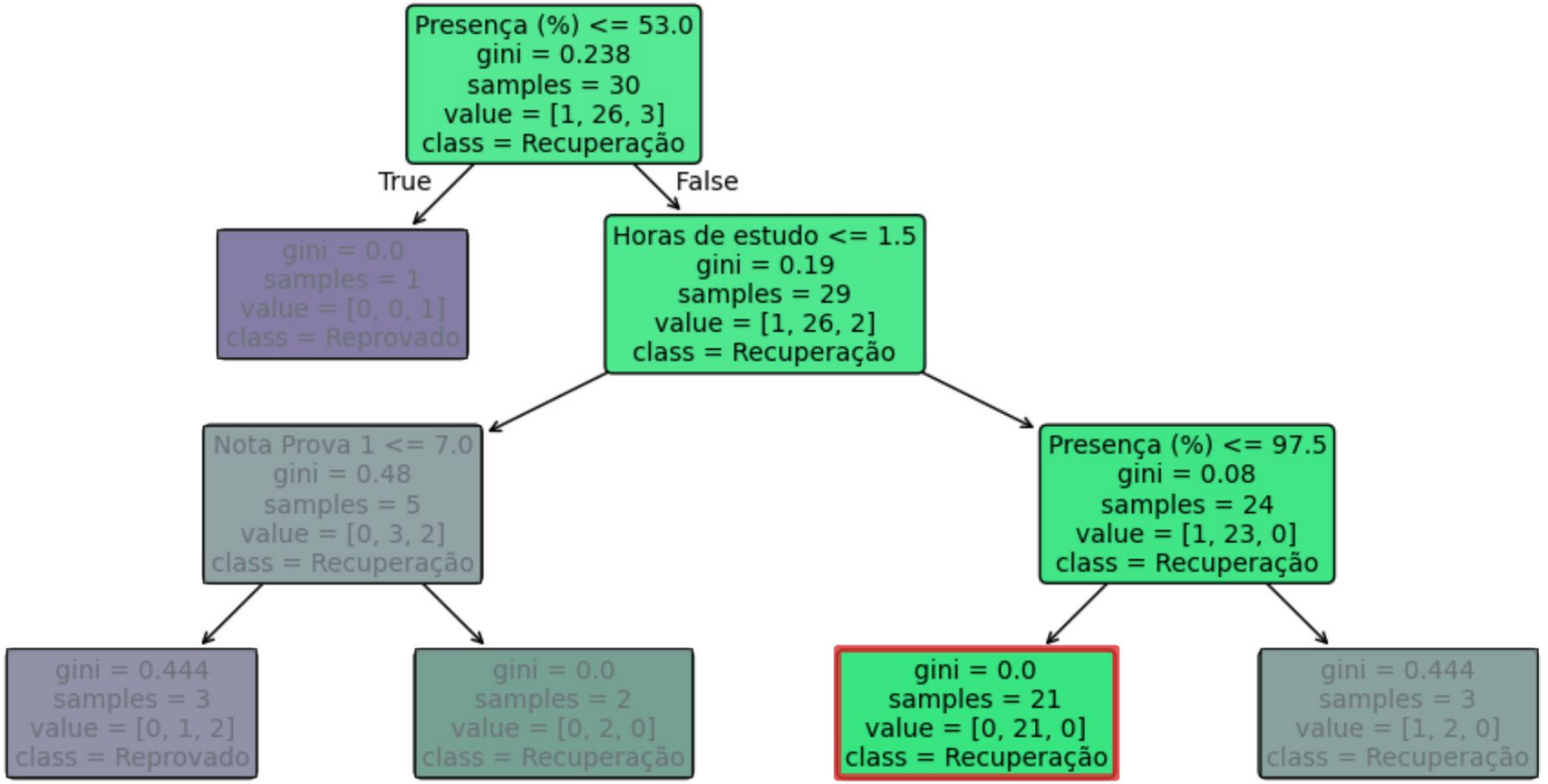
- O modelo escolhe perguntas que **melhor separam as classes** (usando medidas como **Gini** ou **Entropia**).
- Estrutura em forma de **fluxograma**:
  - Nó interno → pergunta (ex.: “Horas de estudo ≥ 4?”)
  - Galhos → caminhos possíveis (sim/não ou intervalos).
  - Folhas → classe final (Aprovado, Reprovado, Recuperação).



ÁRVORE DE DECISÃO

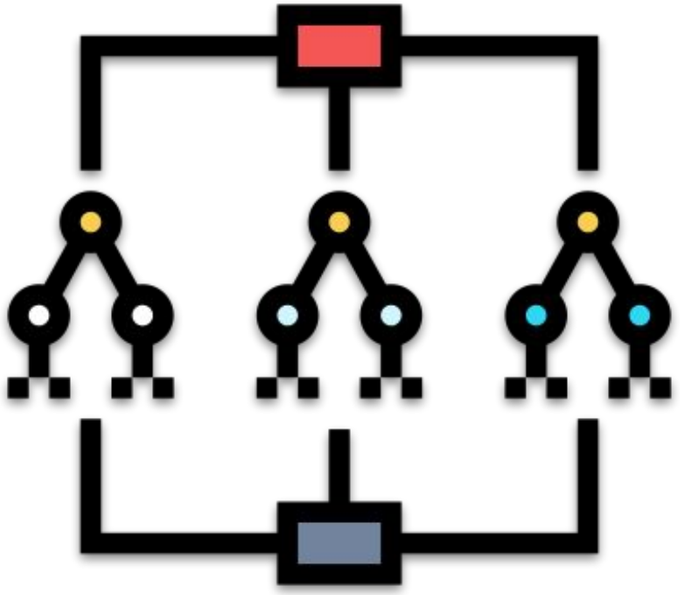


ÁRVORE DE DECISÃO



RANDOM FOREST

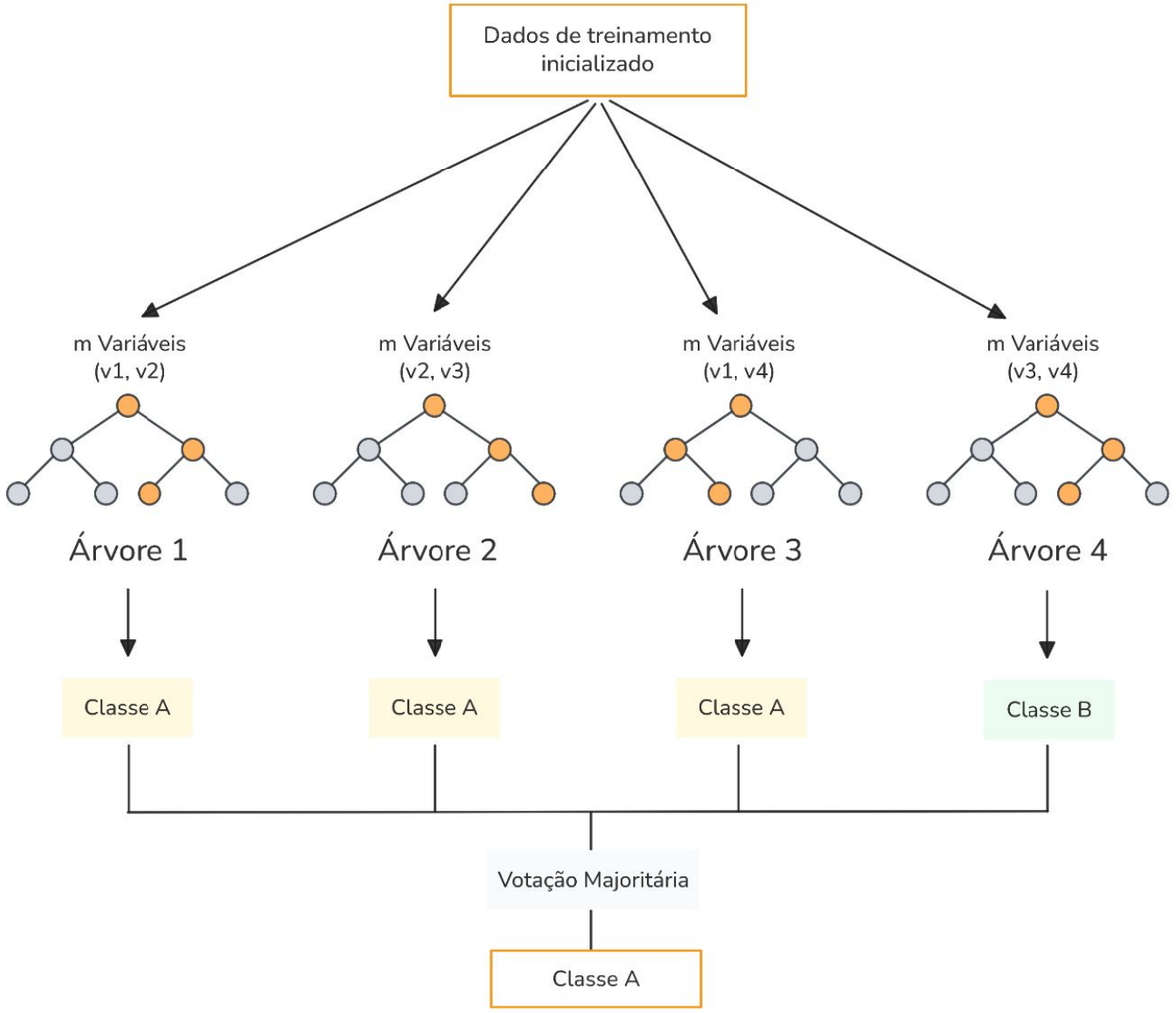
- **Objetivo:** Classificar um novo dado decidindo passo a passo com base em regras, mas melhorando a previsão combinando várias **árvores de decisão**.
- Em vez de usar só uma árvore (que pode ser instável), a Random Forest constrói **muitas árvores** em paralelo.
- Cada árvore é treinada com:
  - Um **subconjunto dos dados** (amostragem com reposição – *bagging*).
  - Um **subconjunto das variáveis** em cada divisão.
- O resultado final é decidido por **votação**:
  - Cada árvore "dá seu palpite" (classe).
  - A classe mais votada é a escolhida.



RANDOM FOREST

Dados de treinamento
Tamanho do Dataset, N=6
N° de variáveis, F=4

v1	v2	v3	v4	y
2.1	0	400	-9	A
3.0	1	890	-42	B
2.2	1	929	0	B
4.0	0	324	-23	A
3.5	1	333	-15	A
6.0	0	215	-9	A



IEEE PES Power & Energy Society®

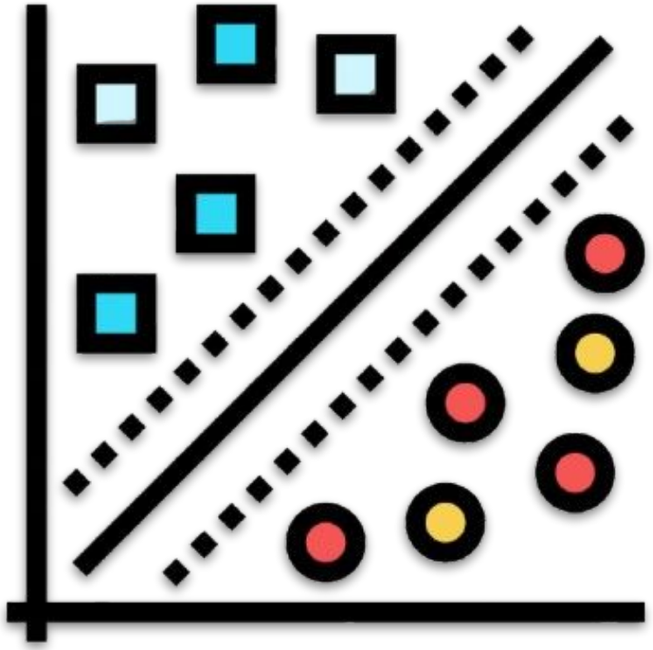
C2AI

CAECOMP

IEEE COMPUTER SOCIETY

SUPPORT VECTOR MACHINE

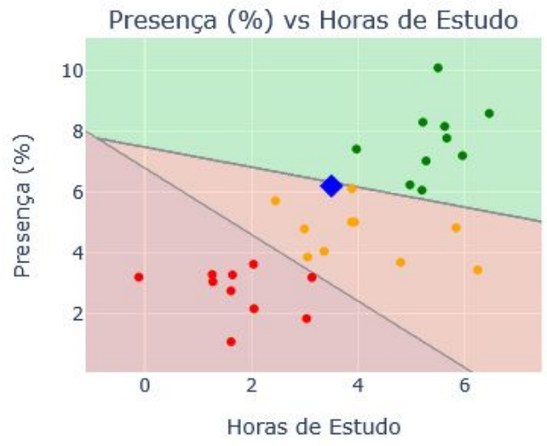
- **Objetivo:** encontrando **fronteiras de decisão** que melhor separam as classes no espaço de variáveis
- O SVM tenta encontrar **hiperplanos** (linhas, em 2D; planos, em 3D; ou hiperplanos em dimensões maiores) que separem as classes.
- Ele não apenas separa, mas busca a **margem máxima**, ou seja, a fronteira que deixa a maior “folga” possível entre os pontos mais próximos de classes diferentes.
- Esses pontos mais próximos da fronteira são chamados de **vetores de suporte** (daí o nome do modelo).



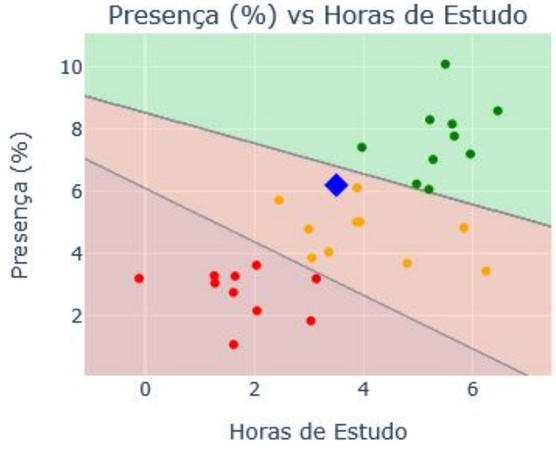
SUPPORT VECTOR MACHINE

Principais Hiperparâmetros

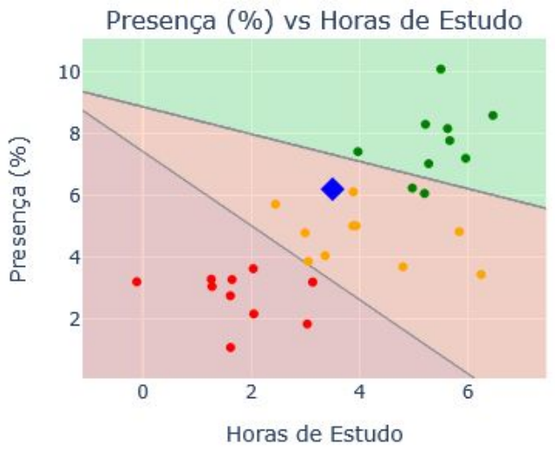
- **C** (Regularização): controla a rigidez da separação.
  a. **C grande** → separação mais rígida, menos tolerância a erros
  b. **C pequeno** → mais tolerância a pontos mal classificados.



C=1.0



C=0.1

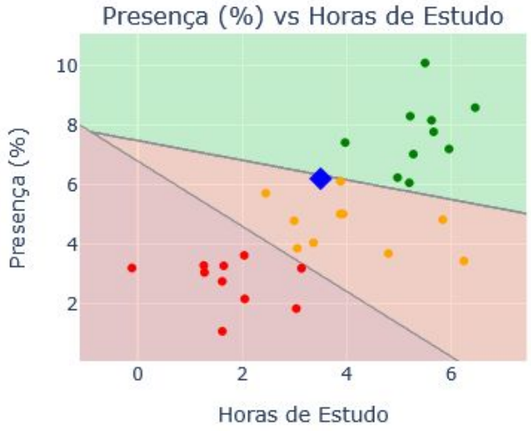


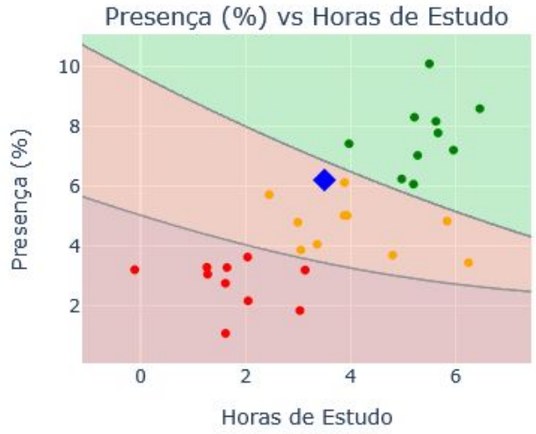
C=0.001

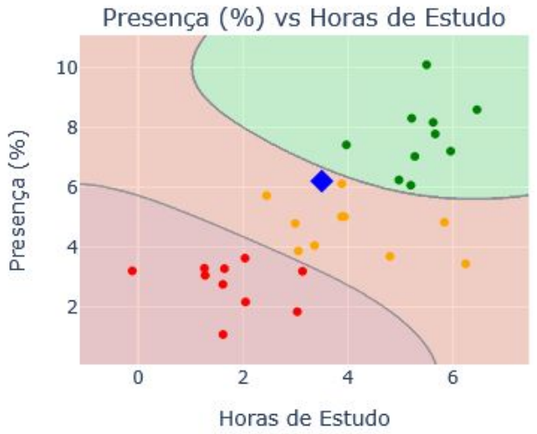
SUPPORT VECTOR MACHINE

Principais Hiperparâmetros

- **Kernel**: função que permite criar fronteiras não lineares.
  a. **Linear**: usado quando os dados já são separáveis com linhas/planos.
  b. **Polinomial**: permite fronteiras curvas mais complexas.
  c. **RBF (radial basis function)**: muito usado, cria fronteiras arredondadas ou “bolhas” em torno das classes.







IEEE PES Power & Energy Society®

C2AI

CAECOMP

IEEE COMPUTER SOCIETY

VANTAGENS E DESVANTAGENS

Modelo	Objetivo principal	Vantagens	Desvantagens	Quando usar
Regressão Logística	Prever a probabilidade de uma classe.	Simples, rápido, fácil de interpretar; Fornece probabilidades.	Assume relação linear; não funciona bem com dados complexos e não lineares.	Quando os dados têm relação aproximadamente linear e interpretabilidade é importante.
KNN (K-Nearest Neighbors)	Classificar baseado nos vizinhos mais próximos no espaço de variáveis.	Simples de entender, não exige treinamento, funciona bem com fronteiras complexas.	Lento com muitos dados; sensível a escala das variáveis e ruído; escolha de K é crítica.	Bases pequenas/médias.
Árvore de Decisão	Criar divisões (regras) em variáveis para classificar exemplos.	Fácil de interpretar (fluxograma), rápido, lida com variáveis categóricas e numéricas.	Pode superajustar (overfitting); instável com pequenas mudanças nos dados.	Quando interpretabilidade é essencial (ex.: explicar decisões para gestores).
Random Forest	Conjunto de várias árvores que votam no resultado final.	Reduz overfitting, alta acurácia, robusto a ruído e outliers.	Menos interpretável, pode ser lento em grandes bases.	Quando precisa de boa performance sem perder robustez, mesmo sem explicabilidade.
SVM (Support Vector Machine)	Encontrar hiperplanos que separam classes com a maior margem possível.	Muito bom em alta dimensão; funciona bem com fronteiras complexas; forte em bases pequenas.	Difícil de ajustar parâmetros (kernel, C, γ); pouco interpretável; lento em bases grandes.	Quando os dados não são lineares e há fronteiras complexas entre classes.

IEEE PES Power & Energy Society®

C2AI

CAECOMP

IEEE COMPUTER SOCIETY

HANDS-ON

https://github.com/conect2ai/Classification-Bootcamp/