

*Instituto Nacional de
Telecomunicações - INATEL*

EDGE AI

IA aplicada a dispositivos de borda

O QUE É IA?

A Inteligência Artificial (IA) é o campo da ciência da computação dedicado a criar sistemas capazes de realizar tarefas que normalmente exigiriam inteligência humana, como perceber o ambiente, aprender com dados, raciocinar e tomar decisões. Algumas áreas são:

- Redes Neurais Artificiais;
- Machine Learning (Aprendizado de Máquina);
- Visão Computacional;
- Processamento de Linguagem Natural (PLN);
- Robótica Inteligente;
- Processamento de Áudio e Voz;



APLICAÇÕES EM IOT

- **Automação:** Controle de qualidade → Câmeras em linhas de produção identificando defeito em peças.
- **Saúde:** Monitoramentos de pacientes com IA ajudando a reconhecer problemas
- **Agricultura:** Monitoramento de plantações, automação de equipamentos...
- **Cidades Inteligentes:** Gestão de tráfego, coleta de lixo com a IA planejando a rota mais eficiente para os caminhões de lixo, garantindo que as lixeiras cheias sejam esvaziadas sem que os caminhões percam tempo em lixeiras vazias.



VISÃO COMPUTACIONAL

A Visão Computacional é uma área da Inteligência Artificial que permite que computadores e dispositivos interpretem e compreendam o mundo visual, analisando imagens e vídeos da mesma forma que os humanos fazem. Por exemplo, ajuda a:

- Reconhecer rostos; como quando desbloqueia tablets usando seu rosto.
- Detectar carros na rua, para veículos autônomos.
- Auxiliar médicos a encontrar doenças em exames.

Tudo isso acontece porque o computador usa redes neurais — como cérebros artificiais — para aprender a “ler” as figuras.

PROCESSAMENTO DE BORDA

Forma onde os dados são processados onde são gerados, ao inves de mandar para nuvem.

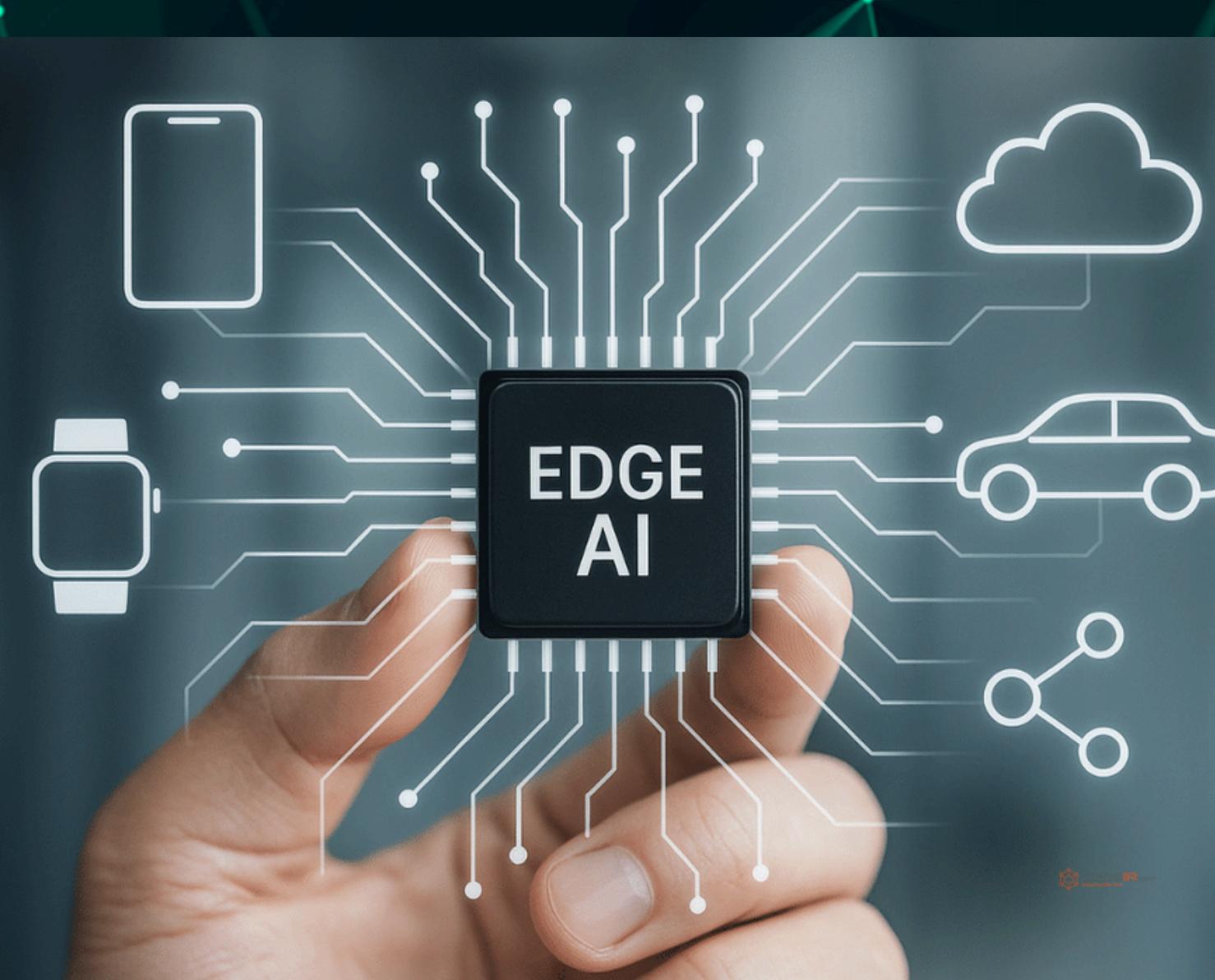
Isso deixa os sistemas mais rápidos, confiáveis e menos dependente de internet.

Exemplo :

- Cameras inteligentes
- Carros autonomos
- Sensores



EDGE AI

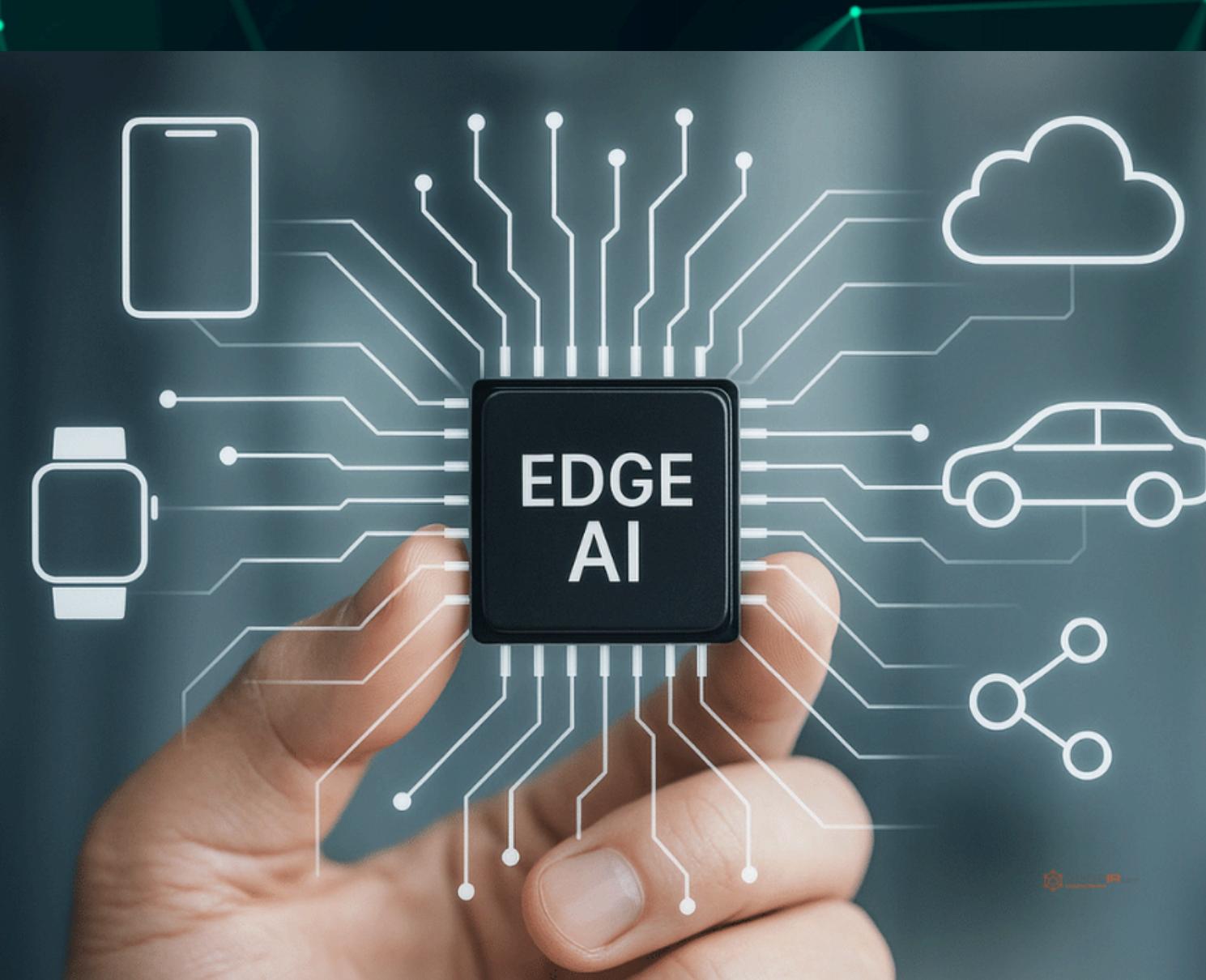


Quando a inteligencia artificial roda direto nos dispositivos, onde processamos os dados sao gerados e processados sem nivel.

Colocando o cerebro como cameras, sensores e placas em tempo real.

Para esse processo precisamos quantizar uma rede, para deixar mais rapido o processamento, mais seguro e consegue funcionar com ou sem internet.

EDGE AI



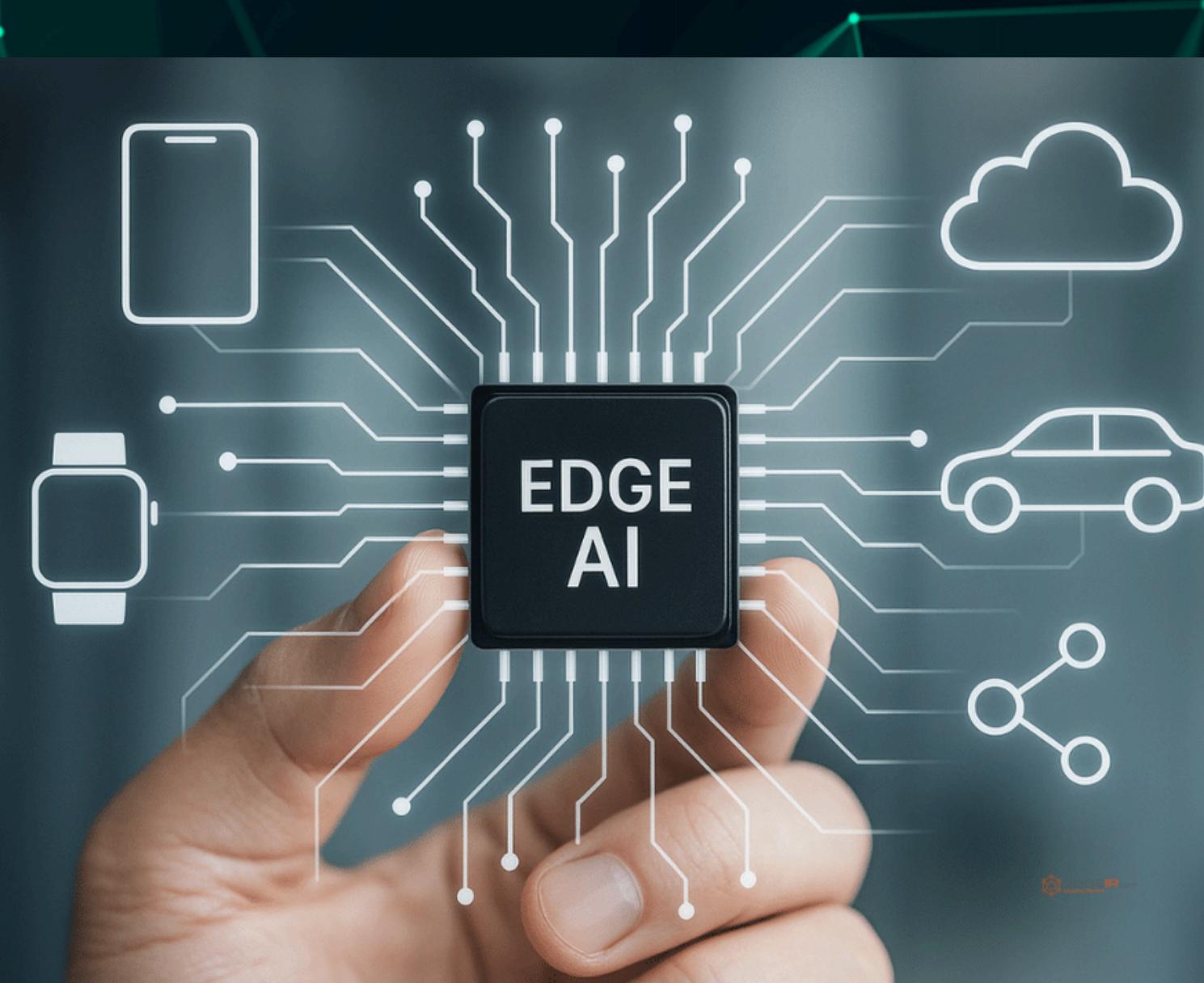
IA NA NUVEM X IA NA BORDA

- Nuvem
- Envia dados brutos → processa na nuvem → devolve resposta
 - Muito poder de processamento
- - Latência alta, depende de internet, mais custo de envio de dados
- Borda (Edge AI)
- Processa local → envia só resultado ou resumo
 - Baixa latência, funciona offline, mais privacidade
- - Limitação de memória, CPU, energia

EDGE AI

LIMITAÇÕES

- Hardware limitado: pouco RAM, pouca flash, CPU fraca.
- Consumo de energia: muita inferência pode drenar bateria.
- Atualização de modelos: como mandar firmware ou modelos novos para vários dispositivos.
- Segurança: dispositivo físico mais exposto (pode ser roubado, aberto, etc.).



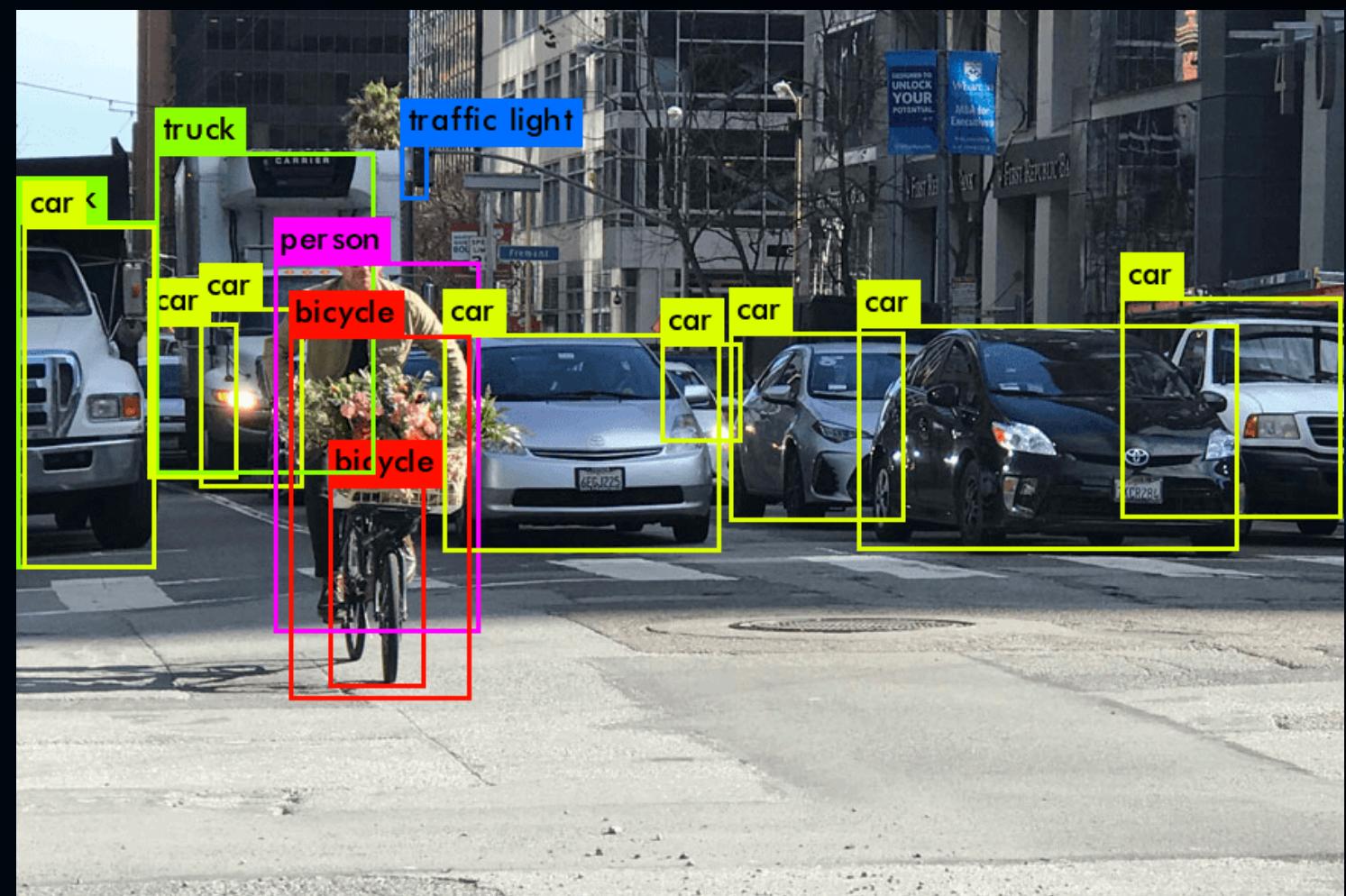
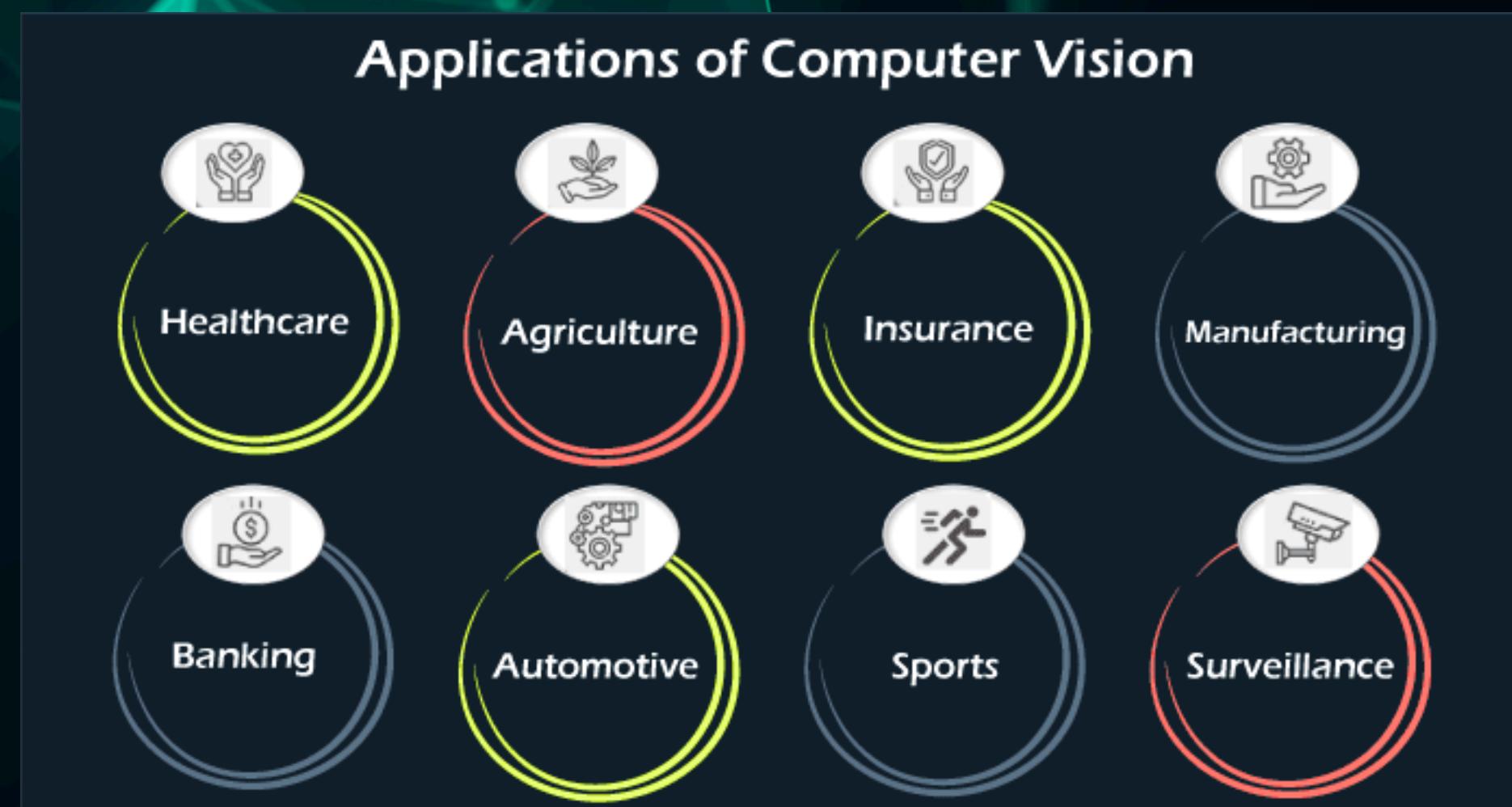
QUANTIZAÇÃO

Para embarcar modelos de IA em dispositivos IoT de baixo custo, como a Raspberry Pi Pico, Pi Pico 2 ou ESP32, é necessário reduzir o tamanho do modelo treinado, originalmente em FLOAT32, sem comprometer significativamente sua precisão.

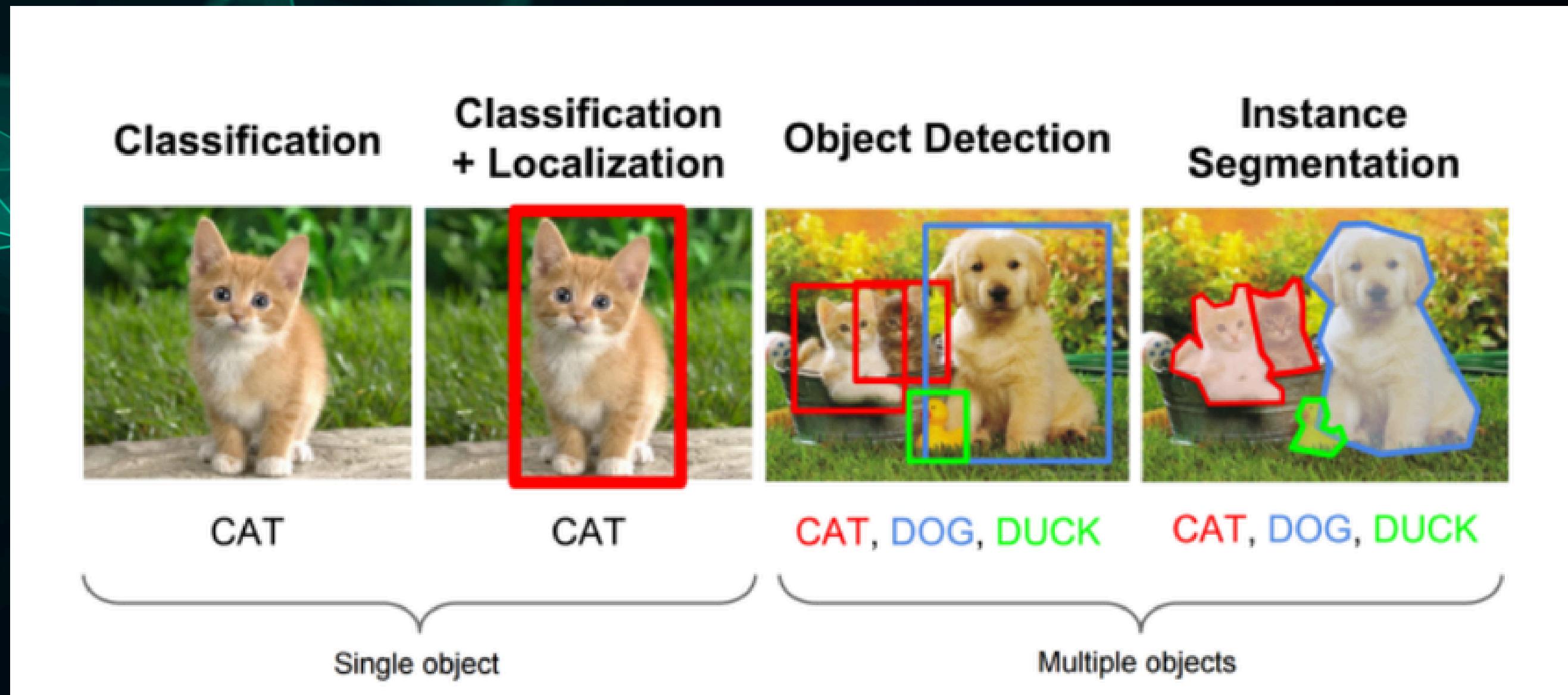
Para isso, aplicamos a quantização para INT8, que diminui o consumo de memória e o tamanho do modelo, permitindo que ele seja executado diretamente no microcontrolador.



ONDE PODEMOS APLICAR A VISÃO COMPUTACIONAL?



VISÃO COMPUTACIONAL



VISÃO COMPUTACIONAL

Classificação de Imagens

→ A classificação de imagens consiste em analisar seu conteúdo visual para atribuir-lhe uma ou mais categorias (rótulos). O processo decide a qual classe a imagem pertence ao analisá-la como um todo.

Detecção de objetos

→ A detecção de objetos é o processo de localizar e atribuir uma ou mais categorias (rótulos) a múltiplos objetos dentro de uma única imagem, com base em seu conteúdo visual. Por ser capaz de identificar e classificar vários itens simultaneamente, ela é considerada um avanço em relação à classificação de imagens.

Segmentação

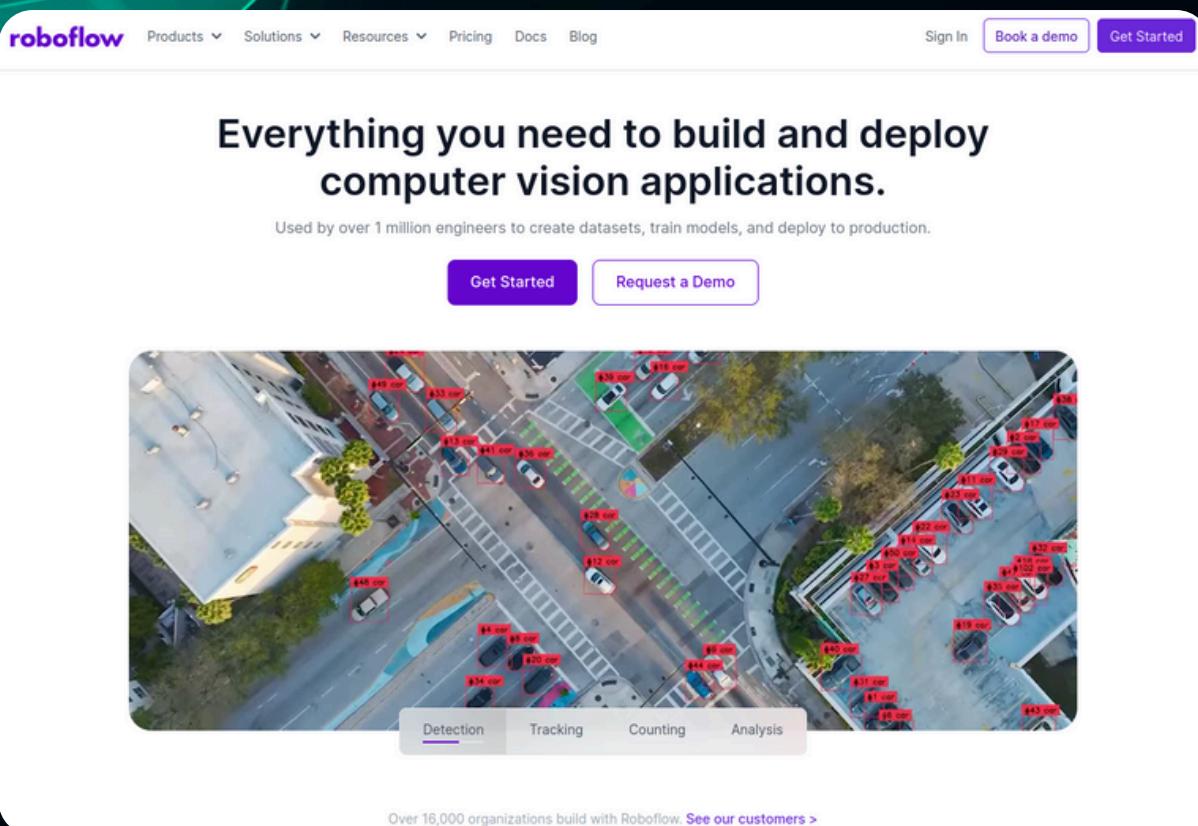
→ A segmentação é a tarefa de identificar e separar diferentes regiões ou objetos dentro de uma imagem. Ao contrário da classificação, que dá um rótulo para toda a imagem, a segmentação fornece informações detalhadas sobre cada pixel.

ETAPAS

1. Definição do problema e coleta de dados
2. Pré-processamento de dados;
3. Escolha do modelo e arquitetura;
4. Configuração do treinamento;
5. Treinamento do modelo
6. Avaliação do modelo;*
7. Otimização para dispositivos IoT;
8. Implementação no dispositivo;

ROBOFLOW

Facilita o treinamento de modelos de IA com ferramentas para anotação, preparo de dados e treinamento.



- Anotação de imagens: Criação fácil de datasets rotulados para treinamento.
- Aumento de dados: Geração automática de variações das imagens para melhorar o desempenho do modelo.
- Treinamento e exportação: Suporte a modelos como YOLO e exportação para TensorFlow, PyTorch, TFLite e ONNX.
- Integração com IA: Compatível com bibliotecas e frameworks populares de deep learning.

TREINAMENTO DA REDE

Código para o algoritmo de treinamento ht

AVALIAÇÃO DO MÓDELO

01. Acurácia (Accuracy)

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

Mede a proporção total de acertos do modelo.

02. Precisão (Precision)

$$\text{Precisão} = \frac{VP}{VP + FP}$$

Mede o quanto das previsões positivas estão corretas.

03. Revocação (Recall)

$$\text{Recall} = \frac{VP}{VP + FN}$$

Mede quantos dos casos positivos reais o modelo conseguiu detectar.

04. F1-Score

$$F1 = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

É a média harmônica entre precisão e revocação.

05. Matriz Confusão

Mostra como o modelo classificou cada categoria, permitindo visualizar os acertos e erros.

QUANTIZAÇÃO

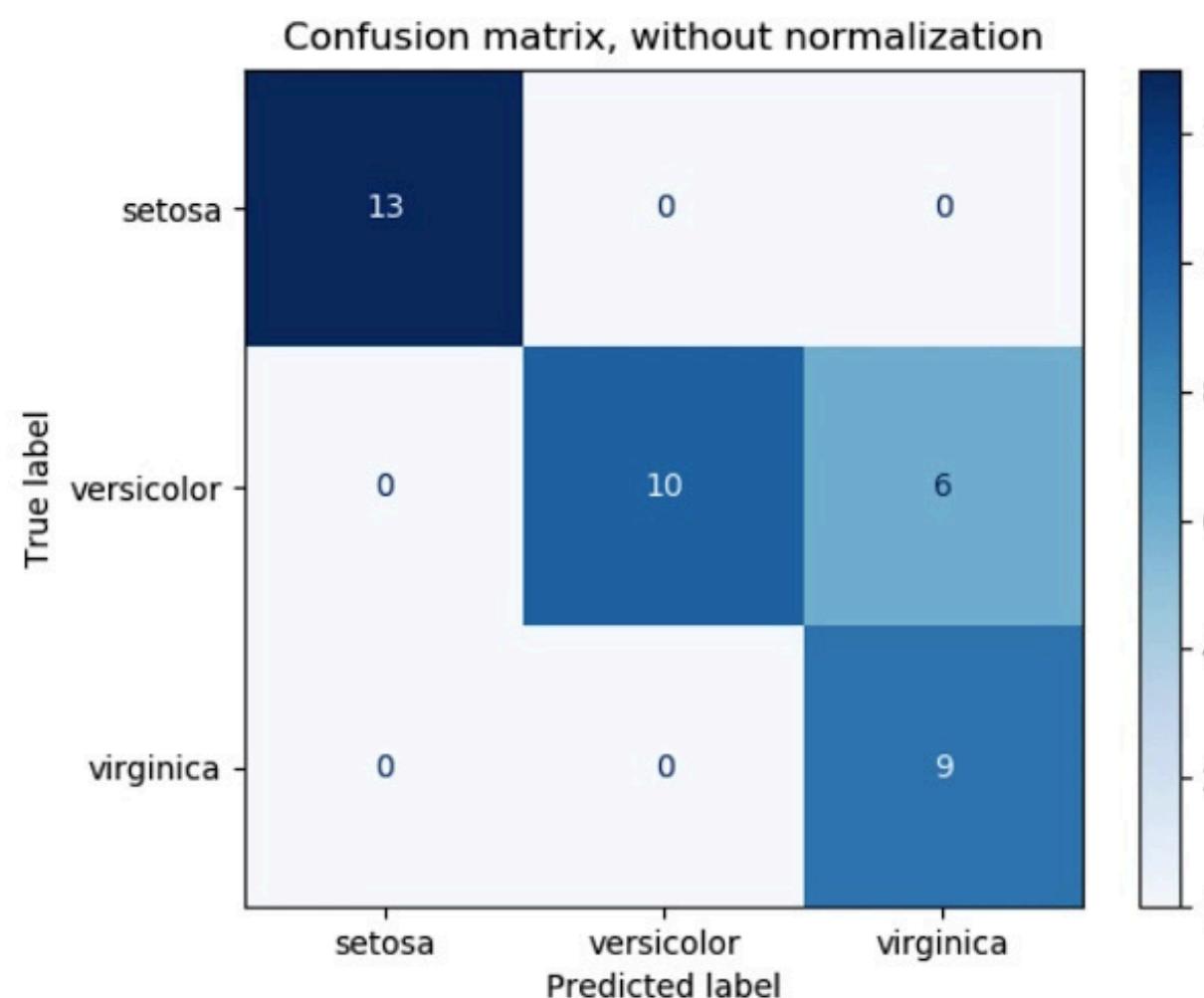
FLOAT32 x INT8

- FLOAT32: números em ponto flutuante de 32 bits. Têm alta precisão, mas ocupam mais memória e deixam o modelo mais pesado.
- INT8: números inteiros de 8 bits. Têm menos precisão, porém o modelo fica até 4x menor, consumindo menos memória e energia.
- Na quantização, convertemos os pesos do modelo de FLOAT32 para INT8, mantendo a precisão aceitável, mas tornando a IA leve o suficiente para rodar em microcontroladores e dispositivos de borda.



MATRIZ CONFUSÃO

Exemplo



DEMONSTRAÇÃO