

Modelos de ML: Tipos, escolhas e estratégias



Plataforma completa de aprendizado
contínuo em programação.

#BoostingPeople

rocketseat.com.br

Todos os direitos reservados © Rocketseat S.A.

O que é um modelo de ML?

Um modelo de ML aprende padrões a partir de dados para fazer previsões ou descobrir relações.

Diferença entre IA (Inteligência Artificial) e ML (Machine Learning)

Inteligência Artificial (IA) = o campo amplo da ciência da computação que busca criar máquinas que imitam comportamentos inteligentes humanos.

Machine Learning (ML) = uma subárea da IA que ensina computadores a aprender com dados.

🎯 Em resumo:

IA = objetivo geral (simular inteligência)

ML = uma das formas de atingir esse objetivo (via aprendizado com dados)

Como funciona?

Etapa	O que acontece
Entrada	Você dá dados (X) + resposta (y)
Aprendizado	O modelo encontra padrões
Avaliação	Você testa se ele aprendeu bem
Predição	Ele faz previsões com base no padrão

Tipos de tarefas em ML

Tipo	O que é?	Quando usar?
Regressão	Prever valores numéricos contínuos com base em variáveis explicativas.	Quando a saída que você quer prever é um número real (R\$, tempo, quantidade, nota, etc).
Classificação	Prever categorias ou classes, geralmente como "sim/não", "baixo/médio/alto", etc.	Quando a saída esperada não é um número, mas sim uma decisão ou rótulo (ex: fraudulento ou não)
Clusterização	Agrupar dados semelhantes sem ter um rótulo pré-definido. O algoritmo descobre padrões escondidos nos dados.	Quando não há resposta certa nos dados (sem y) e o objetivo é explorar, descobrir grupos etc
Redução de dimensionalidade	Transformar um conjunto de dados com muitas variáveis em um conjunto menor, preservando o máximo de informação possível.	Quando você quer facilitar a análise exploratória, melhorar performance de modelos etc

ML supervisionado x não supervisionado

Supervisionado: Você ensina o modelo com respostas corretas.

Não-supervisionado: Você entrega os dados, mas não diz a resposta (o modelo explora padrões)

Aspecto	Supervisionado	Não supervisionado
Tem variável de resposta (y)?	✓ Sim	✗ Não
Objetivo principal	Prever um resultado conhecido	Descobrir estrutura ou padrões ocultos
Exemplo	Prever churn ou vendas	Agrupar perfis de cliente
Exemplos de modelo	Regressão, Random Forest	K-Means, PCA

Exemplos de supervisionado

Regressão Linear

- É um modelo supervisionado que estima valores contínuos com base em uma relação linear entre as variáveis.
- Ex: prever faturamento com base em investimento em marketing.

Regressão Logística

- É um modelo supervisionado usado para classificação binária, que retorna a probabilidade de um evento acontecer.
- Ex: prever se um cliente vai ou não cancelar o serviço.

Exemplos de supervisionado

Random Forest

- É um modelo que junta várias árvores de decisão para tomar uma decisão mais precisa e confiável, como se fosse uma votação em grupo.
- Funciona bem para regressão e classificação, mesmo com dados ruidosos.

XGBoost

- É um modelo que aprende em etapas, corrigindo os erros anteriores a cada rodada, o que o torna muito poderoso para fazer previsões com alta precisão.

Exemplos de não supervisionado

K-Means

- É um modelo que divide os dados em grupos parecidos, colocando juntos os que estão mais próximos entre si, como se formassem “bolhas” de comportamento.

DBSCAN

- É um modelo que encontra grupos com base na densidade dos dados e ainda consegue detectar pontos fora do padrão, sem precisar saber quantos grupos existem.

Exemplos de não supervisionado

▼ PCA (Principal Component Analysis)

- É uma técnica que resume várias variáveis em poucas, mantendo o que mais importa, para facilitar a análise e visualização dos dados.

Regressão vs Classificação

CLASSIFICAÇÃO

Prediz categoria

X

REGRESSÃO

Prediz valores
contínuos

Exemplos de regressão

1. Previsão de Faturamento

▶ Entrada

- Dados históricos de vendas
- Investimento em marketing
- Quantidade de leads por canal
- Sazonalidade (mês, dia da semana)
- Região ou segmento do cliente

🧠 Aprendizado

O modelo aprende a identificar como cada variável influencia o faturamento.

Ex: "Para cada R\$ 1.000 a mais em mídia paga, a receita sobe R\$ 3.500 em média."

Exemplos de regressão

2. Previsão de Tempo de Atendimento

Entrada

- Tipo de solicitação (ex: cancelamento, dúvida técnica)
- Canal de entrada (telefone, chat, e-mail)
- Turno ou horário
- Nível de experiência do atendente
- Tempo médio histórico por tipo de demanda

Aprendizado

O modelo aprende padrões do tempo de atendimento com base nas características da solicitação e do contexto.

Ex: "Chamados técnicos no WhatsApp duram 40% menos tempo do que por e-mail."

Exemplos de regressão

★ 3. Previsão de Nota de NPS

▶ Entrada

- Tempo de resposta a chamados anteriores
- Volume de interações com suporte
- Última ação no produto (ex: erro, compra, renovação)
- Tipo de plano ou pacote contratado
- Histórico de notas anteriores (se houver)

🧠 Aprendizado

O modelo aprende a identificar quais fatores influenciam um cliente a dar uma nota baixa ou alta.

Ex: "Clientes com 3+ chamados não resolvidos têm 65% de chance de dar nota < 7."

Exemplos de classificação

1. Vai ou não comprar? (Classificação binária)

▶ Entrada

- Interações com o site ou app
- Origem da campanha (e-mail, rede social, tráfego orgânico)
- Número de visitas anteriores
- Tempo de navegação
- Produtos visualizados ou adicionados ao carrinho

Aprendizado

- O modelo aprende padrões de comportamento que indicam alta ou baixa propensão de compra.
- Ex: “Usuários que adicionam 2+ produtos ao carrinho e voltam no mesmo dia têm 82% de chance de comprar.”

Exemplos de classificação

✉ 2. E-mail é spam ou não? (Classificação binária)

▶ Entrada

- Palavras-chave no assunto e corpo do e-mail
- Frequência de envio do remetente
- Presença de links suspeitos ou anexos
- Histórico de mensagens semelhantes marcadas como spam
- Horário de envio

🧠 Aprendizado

- O modelo aprende a distinguir padrões linguísticos e comportamentais comuns em e-mails de spam.
- Ex: “E-mails com links + palavras como 'grátis', 'promoção', 'clique aqui' têm alta chance de serem spam.”

Exemplos de clusterização

🌱 1. Descobrir segmentos de clientes (clusterização)

▶ Entrada

- Quantidade de compras
- Ticket médio
- Frequência de compras
- Tempo como cliente
- Produtos mais comprados

🧠 Aprendizado

- O algoritmo (ex: K-Means) agrupa os clientes em clusters com base na similaridade entre seus comportamentos.
- Ex: “Esses 3.000 clientes se dividem naturalmente em 4 perfis distintos de consumo.”

Exemplos de clusterização

2. Mapear comportamentos similares (clusterização)

▶ Entrada

- Ações do usuário em um app ou site (cliques, tempo, ordem)
- Tipo de navegação (pular etapas, voltar, repetir ações)
- Dispositivo e horário de uso
- Perfil demográfico (idade, cidade, canal de origem)

Aprendizado

- O modelo (ex: DBSCAN) encontra padrões de comportamento que ocorrem com frequência — e também pode identificar outliers (comportamentos muito fora do comum).
- Ex: “Usuários que visitam o site à noite e clicam 3x no mesmo botão antes de comprar formam um grupo específico.”

Exemplos de redução de dimensão

🕒 1. Compressão de informação (Redução de Dimensionalidade)

▶ Entrada

- Base com muitas variáveis (ex: 50+ colunas)
- Pode incluir comportamento de usuário, métricas operacionais, variáveis financeiras, etc.
- Exemplo real: matriz com 200 produtos x 10.000 clientes

🧠 Aprendizado

- Algoritmos como PCA (Principal Component Analysis) aprendem a comprimir a informação em menos dimensões, preservando o máximo de variância possível.
- Ex: “As 50 colunas podem ser representadas por apenas 2 ou 3 componentes principais, sem perder muito conteúdo.”

Case 1

Uma empresa de crédito quer prever se um cliente vai ou não ficar inadimplente nos próximos 3 meses.

Você tem acesso a dados como:

- Renda mensal
- Histórico de pagamentos
- Idade
- Quantidade de empréstimos ativos
- Score de crédito interno
- Status de inadimplência (sim ou não)

Qual tipo de modelo de machine learning seria mais apropriado para esse caso?

Case 1: Prever inadimplência

✓ Tipo de problema:

- Classificação supervisionada binária
- (Queremos prever se o cliente vai ou não ficar inadimplente → resposta = 0 ou 1)

✓ Modelos sugeridos:

- Regressão Logística → simples, interpretável
- Random Forest Classifier → lida bem com dados mistos e não lineares
- XGBoost Classifier → alta performance, ideal para cenários de risco e volume

Case 2

Você trabalha em uma empresa de e-commerce e precisa prever o valor de vendas da próxima semana com base em dados históricos como:

- Investimento em marketing digital
- Número de visitas ao site
- Número de itens em estoque
- Semana do ano (para captar sazonalidade)
- Vendas da semana anterior

Qual tipo de modelo de machine learning seria mais apropriado para esse caso?

Case 2: Prever vendas

✓ Tipo de problema:

- Regressão supervisionada contínua (Queremos prever um valor numérico — as vendas futuras)

✓ Modelos sugeridos:

- Regressão Linear → simples, interpretável, bom ponto de partida
- Random Forest Regressor → bom para dados não lineares e variáveis mistas
- XGBoost Regressor → alta performance, ótimo para séries temporais com múltiplas variáveis

Case 3

Seu objetivo é descobrir perfis de clientes com comportamentos parecidos, para aplicar estratégias de marketing mais personalizadas. Você recebeu uma base com 10.000 clientes, contendo variáveis como:

- Quantidade de compras nos últimos 6 meses
- Valor médio gasto por compra
- Frequência de compra
- Canais utilizados (app, site, loja física)
- Categoria de produto mais comprada

Qual tipo de modelo de machine learning seria mais apropriado para esse caso?

Case 3: Segmentar clientes

✓ Tipo de problema:

- Aprendizado não supervisionado (clusterização)
- (Não há variável resposta — queremos descobrir padrões ocultos nos dados)

✓ Modelos sugeridos:

- K-Means → quando você assume que os grupos têm formas esféricas e tamanhos razoavelmente equilibrados
- DBSCAN → quando há ruído ou formatos irregulares nos clusters
- Hierarchical Clustering → se quiser visualizar a formação dos grupos como um dendrograma

Conclusão

- Mais importante que o modelo é entender o problema
- Modelos bons com dados ruins dão respostas ruins