**Introdução aos Métodos de Classificação em Ciência de Dados**

**O que é classificação?**

Classificação é uma técnica de **aprendizado supervisionado** usada para prever categorias ou rótulos discretos com base em dados de entrada. O objetivo é treinar um modelo que consiga identificar a **classe correta** de novas observações, com base em padrões aprendidos a partir de dados rotulados. Nas minhas aulas, eu sempre digo que classificação é colocar os objetos em caixinhas.

**Exemplos:**

* Determinar se um e-mail é “spam” ou “não spam”.
* Determinar se uma notícia é “fakenews” ou não.
* Diagnosticar se um paciente tem ou não uma doença com base em exames.
* Classificar o sentimento de uma avaliação como “positiva”, “negativa” ou “neutra”.
* Classificar alunos dado um conjunto de notas.

**Componentes de um Problema de Classificação**

1. Variáveis independentes (features): características usadas para prever o resultado.
2. Variável dependente (target): a categoria ou classe a ser prevista.
3. Conjunto de treino: usado para treinar o modelo.
4. Conjunto de teste: usado para avaliar a performance do modelo.

**Como funciona?**

Isso depende do método, por exemplo, a regressão logística, gera uma probabilidade associada a cada classe. Já o KNN (K-Nearest Neighbours) busca pontos que são similares, as árvores de decisão criam uma estrutura em árvore para dividir os dados com critérios bem simples, a random forests é um conjunto de múltiplas árvores de decisão, já o SVM (Suport Vector Machine) tenta achar um plano de separação que seja o melhor possível. Como podemos ver o “como” pode ser muito distinto de método para método, mas o objetivo é sempre o mesmo, colocar coisas em caixinhas.

**Nossa primeira classificação**

**Iris Dataset**

* **Objetivo**: Classificar espécies de flores (*Setosa*, *Versicolor*, *Virginica*) com base em características como comprimento e largura das pétalas e sépalas.
* **Tipo**: Classificação multiclasse (3 classes).
* **Tamanho**: Pequeno (150 amostras).

Use o arquivo *‘LDS\_Classificacao\_01.ipynb`* para importar o DataSet e aplicar os algoritmos 'Logistic Regression', 'K-Nearest Neighbors' e 'Decision Tree'.

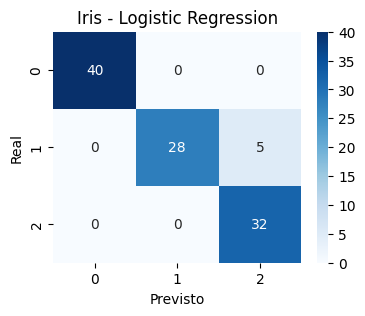
**Avaliando a classificação**

**Classes do Iris Dataset**

As flores do dataset Iris estão divididas em 3 classes (espécies):

| Código | Espécie |
| --- | --- |
| 0 | Setosa |
| 1 | Versicolor |

**Exemplo de Matriz de Confusão para o Iris**



Suponha que você treinou um classificador e obteve a seguinte matriz de confusão:

Essa matriz diz o seguinte:

* Linha 0 (Setosa): Das 40 amostras reais de Setosa, todas foram classificadas corretamente como 0 → excelente desempenho para essa classe.
* Linha 1 (Versicolor):
  + 28 foram corretamente classificadas como 1.
  + 5 foram erradamente classificadas como Virginica (classe 2).
* Linha 2 (Virginica):
  + 32 foram corretamente classificadas como 2

**Como interpretar:**

* Diagonal principal → acertos do modelo.
* Fora da diagonal → erros de classificação, ou seja, confusões entre classes. (por isso o nome matriz de confusão)
* A matriz mostra quais classes são mais fáceis de separar (ex: Setosa) e quais são mais difíceis ou parecidas (Versicolor e Virginica).