

## **Análise de Regressão:**

A análise de regressão é uma abordagem estatística que busca investigar e quantificar a relação entre variáveis. Ela é usada para entender como uma variável dependente (target) está relacionada a uma ou mais variáveis independentes (features).

O objetivo é construir um modelo matemático que represente essa relação, permitindo previsões, inferências, avaliação de impacto e tomada de decisão baseada em dados.

## **Como funciona a regressão:**

- A partir dos dados, estimamos um conjunto de parâmetros (coeficientes).
- Esses parâmetros mostram quanto cada variável independente influencia o resultado.
- O modelo tenta minimizar o erro entre os valores previstos e observados.

### **1. Regressão Linear Simples**

- 1 variável dependente
- 1 variável independente
- O modelo é uma reta:  $y = \beta_0 + \beta_1 x$

### **2. Regressão Linear Múltipla**

- Várias variáveis independentes
- Modelo linear mais complexo
- Permite analisar o impacto de múltiplos fatores simultaneamente

### **3. Regressão Logística**

- Usada para classificação binária
- Estima a probabilidade de uma amostra pertencer a uma classe
- Usa a função sigmoide para transformar a saída linear em probabilidades
- A decisão ocorre com base em um threshold
- Treinada usando entropia cruzada (log loss)
- A função logit (inversa da sigmoide) modela a relação entre chance e probabilidade

#### **4. Regressão Polinomial**

- Modela relações não lineares
- Utiliza termos como  $x^2, x^3x^2, x^3 \dots$
- Aproxima dados curvos mantendo a estrutura da regressão linear

#### **5. Regressão de Poisson**

- Indicada para dados de contagem (ex.: número de ocorrências por intervalo)
- Assume distribuição de Poisson

#### **6. Regressões com Regularização:** Usadas para resolver problemas como multicolinearidade e overfitting:

- Ridge (L2)
  - Penaliza coeficientes grandes
  - Não zera coeficientes
  - Suaviza o modelo
- Lasso (L1)
  - Pode zerar coeficientes
  - Realiza seleção de variáveis
- Elastic Net
  - Combina L1 e L2
  - Bom quando há muitas variáveis correlacionadas

#### **7. Regressão Gamma**

- Usada quando:
  - os dados não seguem normalidade
  - há assimetria positiva
  - o target é contínuo e positivo

## 8. Regressão Beta

- Aplicada quando a variável dependente está restrita ao intervalo (0, 1)
- Muito usada para:
  - proporções
  - taxas
  - conversão
  - probabilidade estimada

## 9. Naive Bayes (para classificação)

- Baseado no Teorema de Bayes
- Assume independência entre as features
- Usa probabilidades condicionais para classificar instâncias
- Extremamente rápido e eficaz em textos (Bag-of-Words)

## 10. Árvore de Decisão

- Uma **Árvore de Decisão** é um algoritmo de aprendizado supervisionado usado para **classificação e regressão**.
- Ela organiza decisões na forma de uma **estrutura hierárquica**, parecida com um fluxograma, onde cada nó representa uma pergunta (condição) e cada ramificação representa uma resposta.
- Usa métricas para medir a "impureza" dos grupos:

### 1. Gini

- Mede o quanto misturado estão os dados
- Quanto menor o Gini, mais “puros” os nós

### 2. Entropia (Information Gain)

- Baseado em teoria da informação
- Quanto maior a informação ganha ao dividir, melhor o split