**Características dos Dados**

* Estruturados

Tabelas de banco de dados, frequências

* Semi Estruturados

JSON, XML e HTML

* Não estruturados

Música, imagem

**Preparação dos Dados**

Para algoritmos que possuem dificuldades com muitos objetos

(+acurácia = +dados = -processamento)

* **Aleatória simples**: exemplos dos dados originais de forma aleatória
* **Estratificada**: exemplos de forma proporcional a quantidade
* **Progressivas**: amostra que aumenta ao longo do desenvolvimento

**Dados desbalanceados provocam**

**Overfitting:** se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados

**Underfitting:** não aprende bem os padrões nos dados de treino, resultando em baixo desempenho tanto nos dados de treino quanto nos dados de teste

**Dados podem ser**

* Incompletos
* Ruidosos (erro ou outliers)
* Inconsistentes
* Redundantes

**Dados incompletos podem ser preenchidos com**

* Eliminando eles
* Preencher automáticamente (“desconhecido”, média dos outros atributos, valor mais provável)
* Preencher manualmente (geralmente inviável)

**Como lidar com Dados ruidosos**

* **Encestamento:** Classificar em faixas, suavizar o ruído pela média
* **Agrupamento:** detectar e remover outliers (n pertencem a um grupo)
* **Regressão**
* **Distância**