

DATA SCIENCE

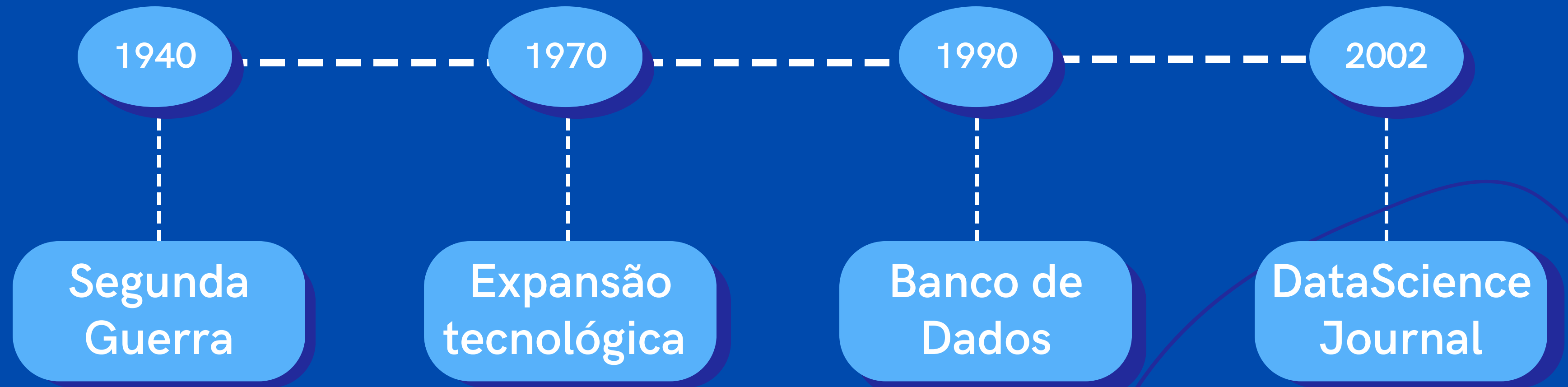


ESTATÍSTICA

- Década de 70
 - Campo Estatístico
 - Criação do termo "Cientista de dados"
 - Limitado poder computacional
- Recursos de organização de dados
 - Bancos de Dados Relacionais
 - Big Data, Business Intelligence ...



VOLTANDO NO TEMPO



O QUE É DATASCIENCE?

PROFISSIONAIS DE DADOS



Engenheiro de dados	Analista de dados	Cientista de dados
<ul style="list-style-type: none">• Coleta e ingestão de dados• Limpeza• Armazenamento• Segurança	<ul style="list-style-type: none">• Coleta e preparação• Exploração de dados• Análise estatística• Insights	<ul style="list-style-type: none">• Análise exploratória• Modelagem de algoritmos• Avaliações e soluções• Implementação de soluções

CIENTISTA DE DADOS: DE OLHO NO FUTURO

CICLO DE VIDA DATASCIENCE



Fonte: Google Imagens

O QUE É UM DATASET?

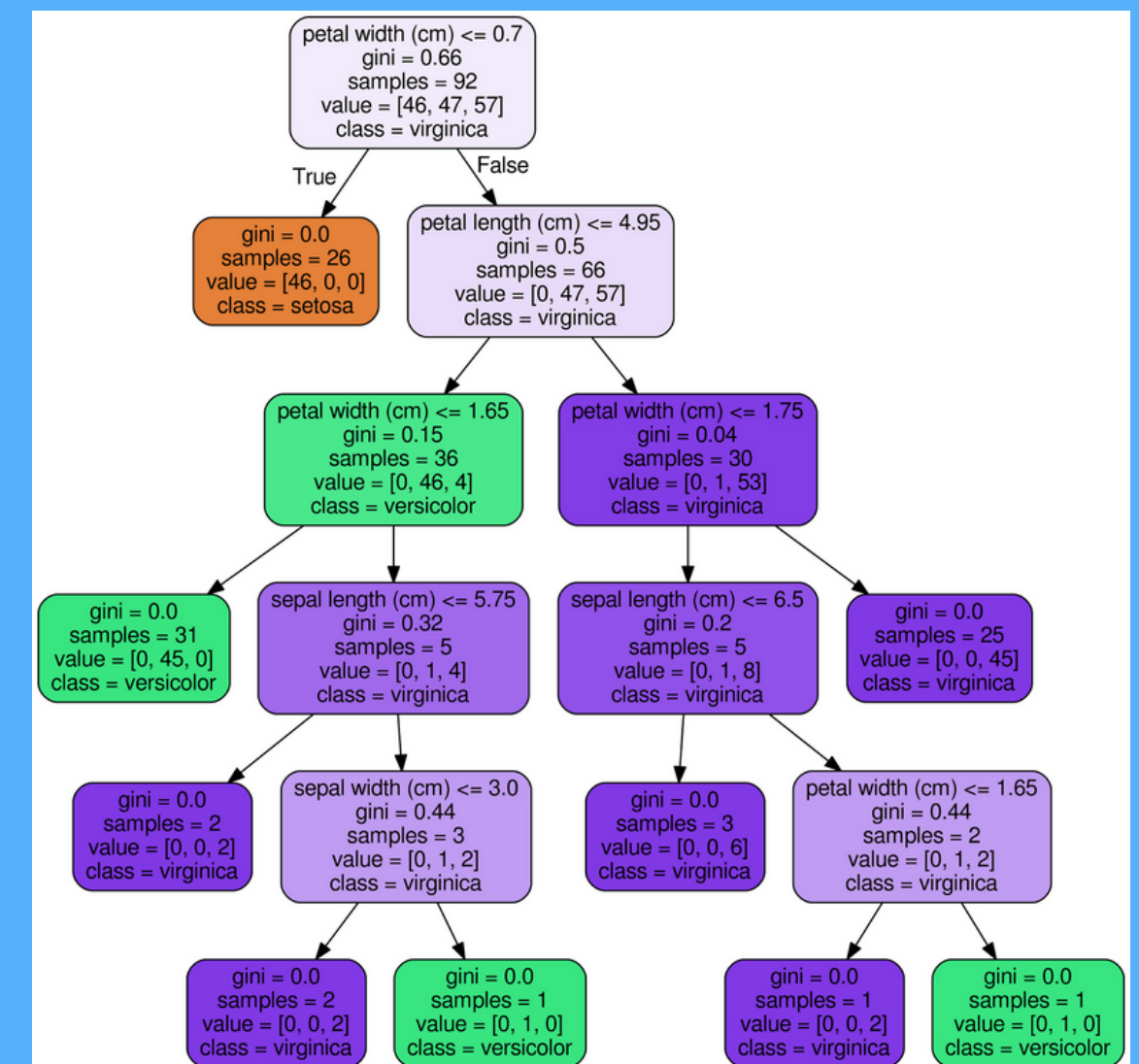
	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows x 6 columns

Fonte: Google Imagens

O QUE SÃO ATRIBUTOS?

- Características em um conjunto de dados
 - Descrevem a amostra em questão
- Tipos de atributos
 - Categóricos
 - Numéricos
 - Temporais
- Descrevem o por que de algo ser o que é



Fonte: Google Images

POR QUE ANALIZAR?

- Insights e compreensão do problema
- Otimização do processo
- Tomada de decisões
- Seleção de atributos
- Validação da solução





EXPLORAÇÃO DOS DADOS

- Compreensão dos dados
- Limpeza dos dados
- Análise descritiva
- Identificação de padrões
- Análise de anomalias
- Análise de relações



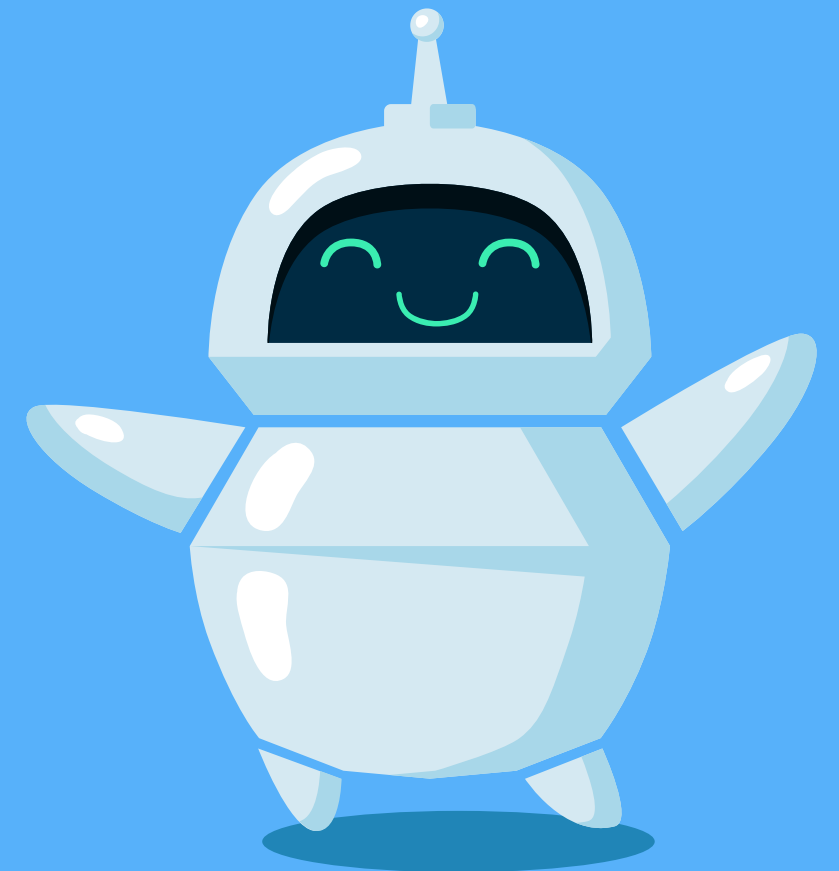
ENGENHARIA DE FEATURES

- Seleção de variáveis
- Criação de novas variáveis
- Transformação de variáveis
- Tratamento de variáveis
- Técnicas de processamento

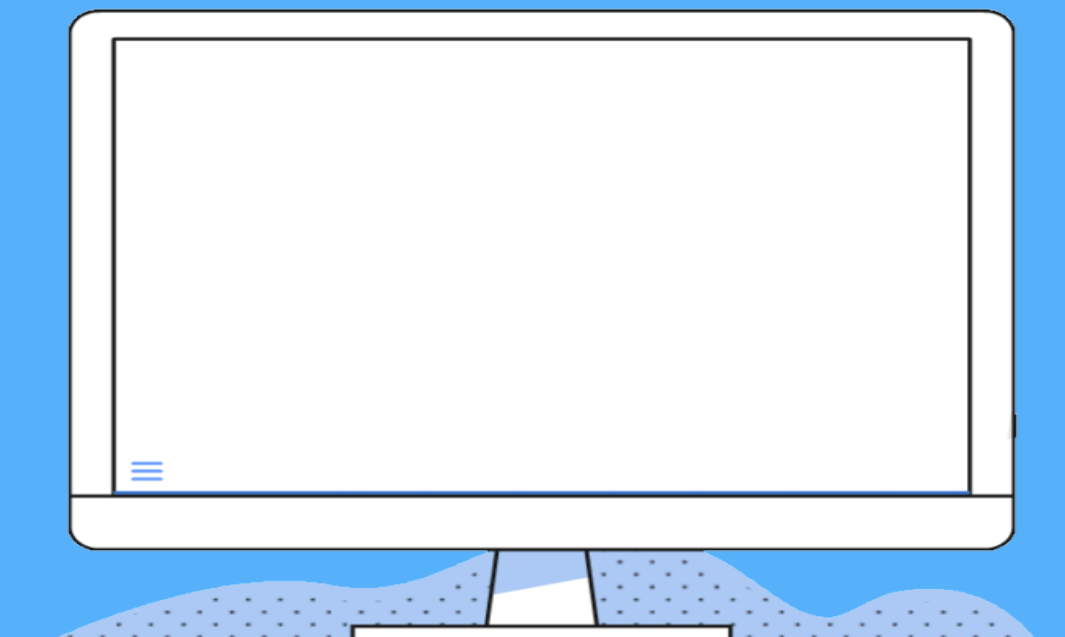


MODELAGEM

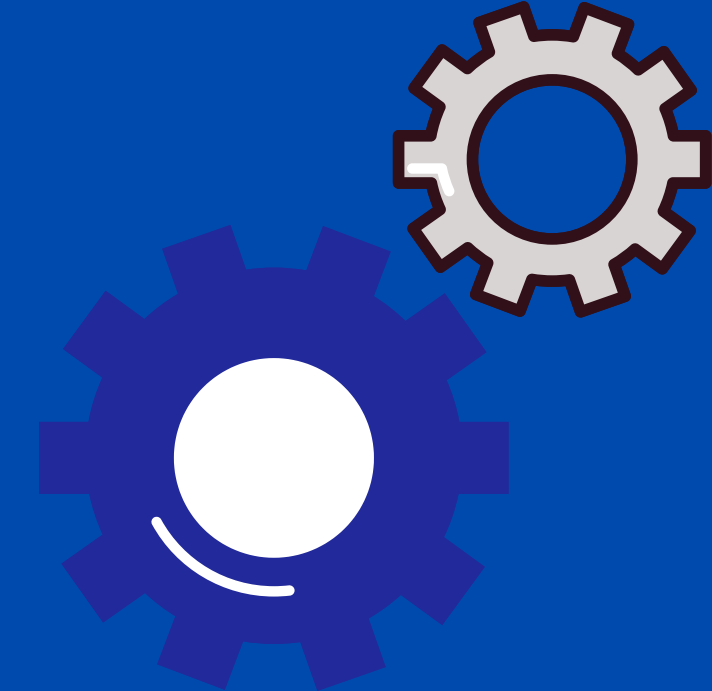
- Definição do problema
 - Classificação X Regressão
- Escolha do algoritmo
- Treinamento do modelo
- Validação do modelo
- Otimização



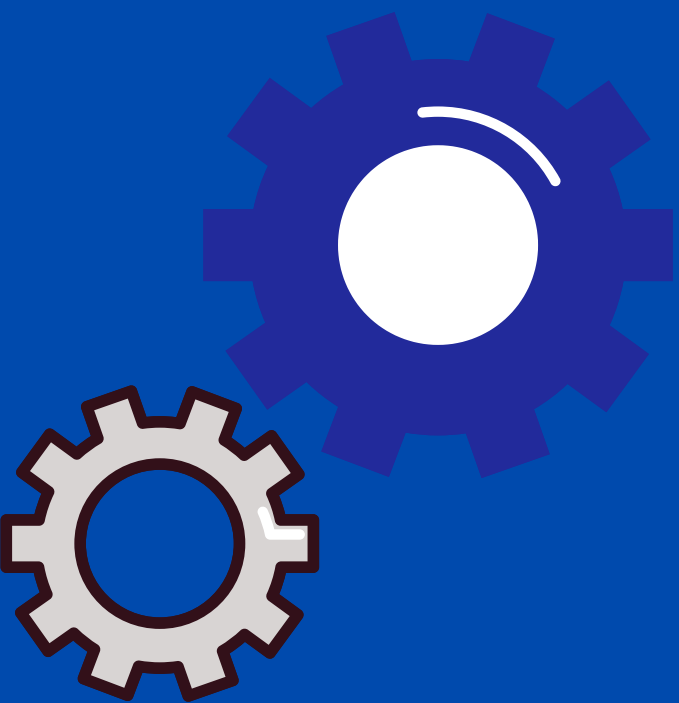
IMPLANTAÇÃO DO MODELO



- Criação da aplicação
- Atualização do Modelo
- Monitoramento contínuo



FERRAMENTAS PARA DATASCIENCE



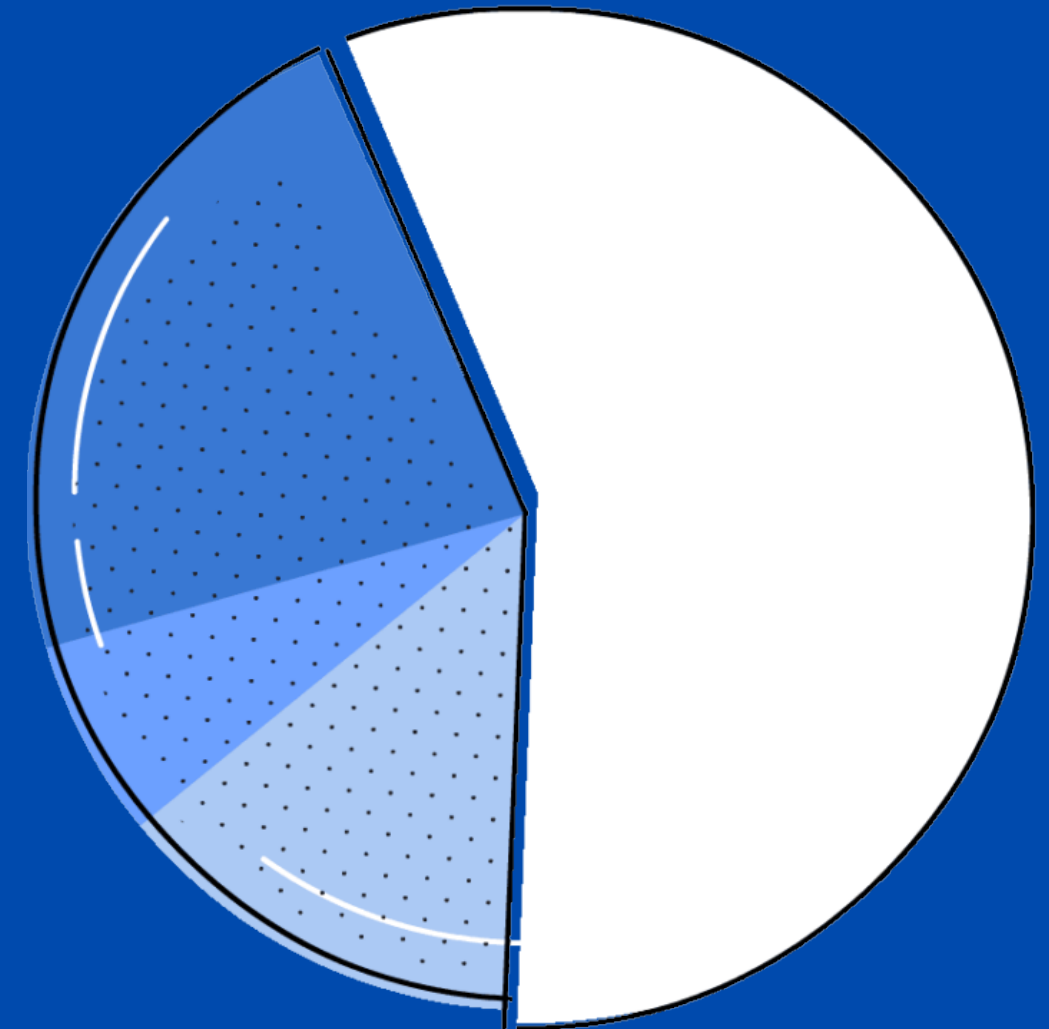
PANDAS

- Manipulação de dados
- Seleção de dados
- Limpeza e transformação de dados
- Análise e exploração do dataframe



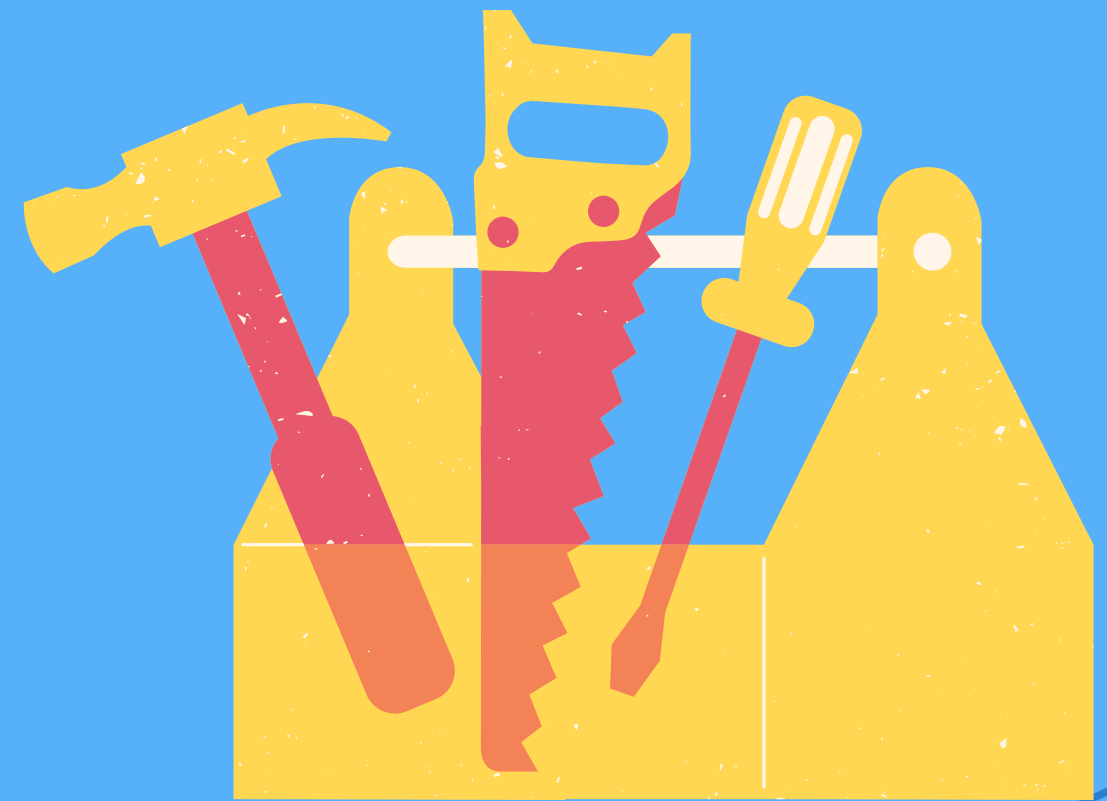
SEABORN

- Visualização de dados
- Gráficos estatísticos
- Personalização
- Suporte a estatística



SCIKIT-LEARN

- Biblioteca para aprendizado de máquina
- Seleção de algoritmos
- Interface consistente
- Avaliação e seleção de modelos



ATIVIDADE 1: EXPLORE UM DATASET

