Análise de dados com Python

e Machine Learning

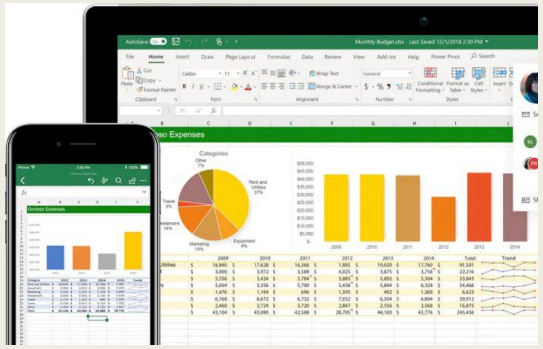
Rafael Silva Pinto

Análise de dados!

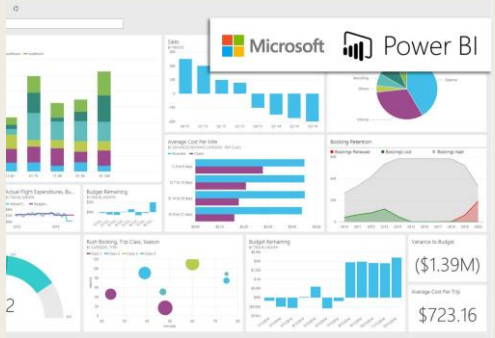
- Busca por redução de custos, aumento de receita e entendimento de problemas.
- Atualmente, uma das áreas de maior destaque dentro das empresas
- Poucos profissionais que realmente sabem aplicar na prática
- Você pode:
  - *Entender o real problema de uma empresa*
  - *Definir setores e produtos que geram maior receita ou menor*
  - *Identificar defeitos*
  - *Direcionar ações*
  - *E muito mais*

Análise de dados!

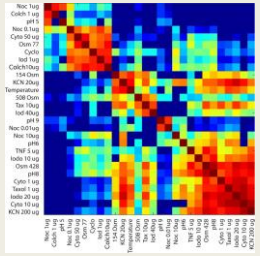
Excel

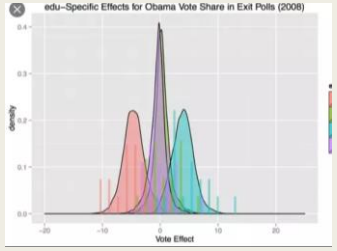


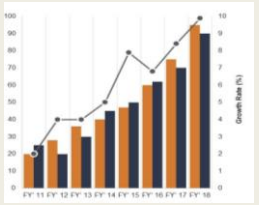
BI

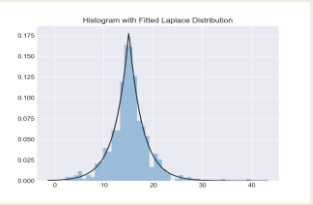


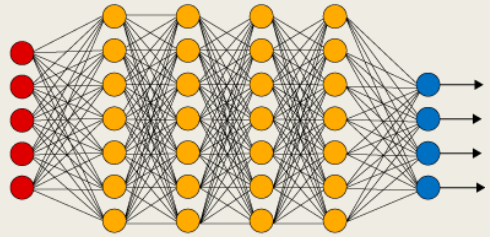
Python + ML







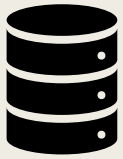




O que você vai aprender

- Instalação do Python
- Começar com Python
- Definição de um caso de estudo
- Leitura de bases
- Tratamento dos dados
- Análise de dados, gráficos e buscar respostas
- Introdução a Machine Learning
- Técnicas de Machine Learning
  - Regressão Logística
  - Árvore de decisão
  - Redes Neurais
  - XGBoost
- Framework de trabalho
- Como continuar e manter a prática

Framework de Trabalho (resumo)



Um problema e dados

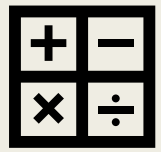
Análises

- Entender o contexto do problema
- Extrair os dados
- Data Cleaning
- Data Fusion
- Feature Engineering
- **Gere Perguntas**
- Interagir e Visualizar

Ache informações e tire ideias

Machine Learning

- Crie sua base com dados os relevantes
- Defina ou crie uma coluna alvo (target)
- Normalize os dados
- Trate os dados categóricos
- Separe base treino e teste
- Teste modelos de ML!



Gere métricas

Gere dinheiro ou salve vidas!

Instalando o Python

- Instalação do Python: https://www.python.org/
- Instalando Bibliotecas: *pip install <biblioteca>*
- IDE Spyder: *pip install spyder*
  no prompt de comando: *spyder*
- Jupyter Lab: *pip install jupyterlab*
  no prompt de comando: *jupyter lab*

Começando com Python

- Lista, tuplas e dicionários
- Laços condicionais: FOR, WHILE e IF
- Functions
- Numpy e Pandas
- Pandas: ler Excel e algumas funções
- Gráficos

Etapas para Análise Dados e Data Science

- Entenda o contexto do problema e o que representam os dados
- Extraia os dados
- Faça a limpeza dos dados (data cleanning)
- Conecte e mescle os dados (data fusion)
- Gere novas informações (feature engineering)
- Interagir, interagir, interagir... visualizar...
- Crie modelos de machine learning

Tipos de Variáveis

```python
# Inteiros
a = 28
print(a)
# Saída: 28

# Ponto flutuante
pi = 3.1415
print(pi)
# Saída: 3.14.15

# String
name = 'Alexsandro Felix'
print(name)
# Saída: Alexsandro Felix

# Boolean
b = True
print(b)
# Saída: True
```

type(var)

Caso de estudo

- Bases do caso de estudo
- Leitura das bases com Python
- Análise preliminar das bases

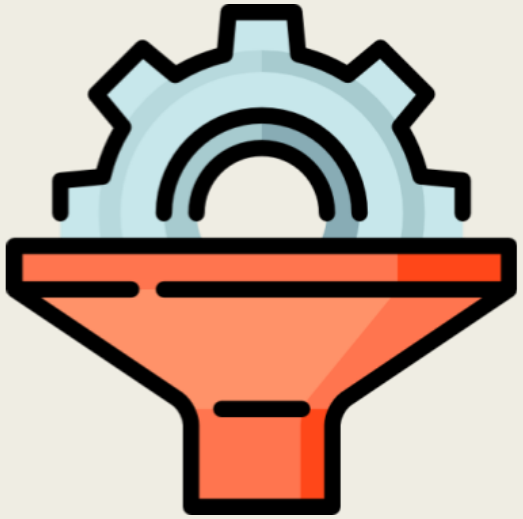
Data Cleanning

- Lidar com os dados Nulos
- Busca por Outliers e tratativas
- Busca por inconsistência entre bases (ex.: alguma venda para cliente inexistente?)
- Busca por dados duplicados
- Tratativas com formato de dados (ex.: datas)
- Identificação de chave primária
- Normalizar?

Juntar bases (Data Fusion)

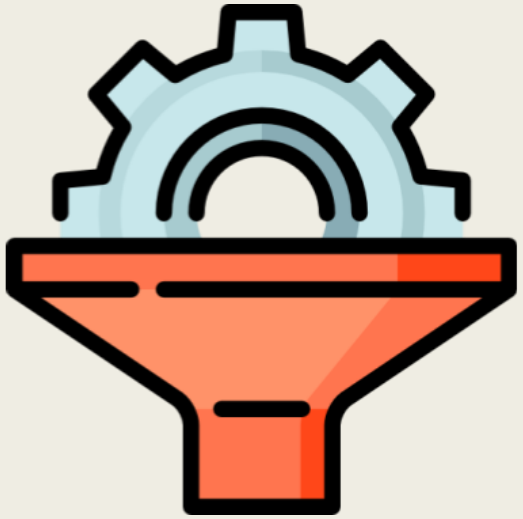
- Join
- Dados nulos e zerados

Criando novos dados (Feature Engineering)



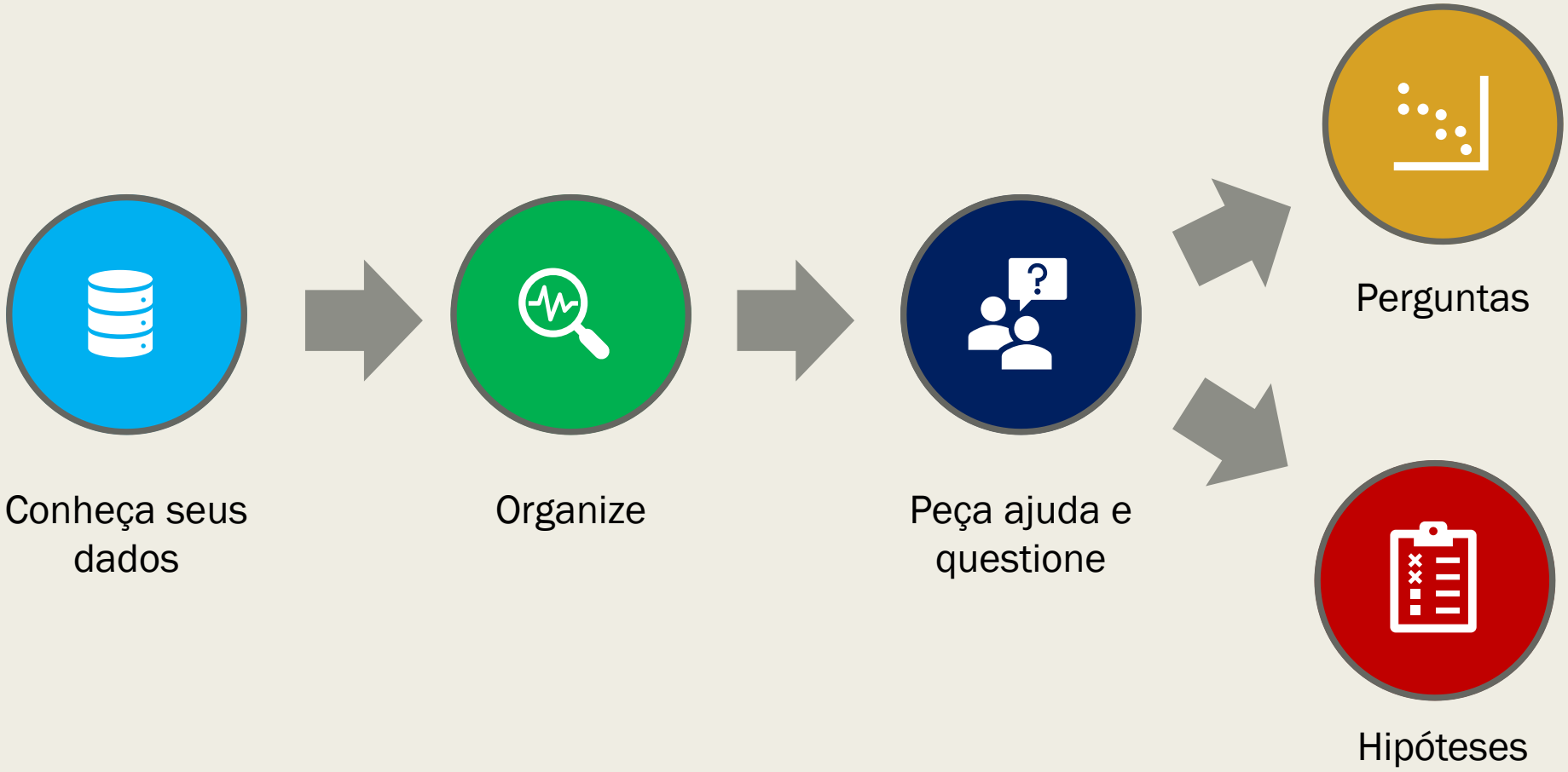
- **Faça perguntas que você gostaria de responder!**
  - *Quais lojas mais vendem?*
  - *Quais produtos mais vendem?*
  - *Quais lojas geram maior receita?*
  - *Quais produtos geram maior receita?*
  - *Existe algum cliente que mais realiza compras?*
  - *Existe alguma relação entre loja e cliente?*
  - *Qual o tempo médio entre compra e pagamento?*
  - *Existe alguma loja em que esse tempo é menor? E produto?*
  - *Qual produto mais gera inadimplência?*
  - *Qual loja tem mais inadimplências?*
  - *Existe alguma relação entre idade e inadimplência?*
  - *É possível prever inadimplência através dos dados idade, cidade e produto?*
- **Gere novos dados ou transforme!**

Criando novos dados (Feature Engineering)



- **Faça perguntas que você gostaria de responder!**
  - *Quais lojas mais vendem?*
  - *Quais produtos mais vendem?*
  - *Quais lojas geram maior receita?*
  - *Quais produtos geram maior receita?*
  - *Existe algum cliente que mais realiza compras?*
  - *Existe alguma relação entre loja e cliente?*
  - *Qual o tempo médio entre compra e pagamento? (**tempo_pg**)*
  - *Existe alguma loja em que esse tempo é menor? E produto?*
  - *Qual produto mais gera inadimplência? (**pg**)*
  - *Qual loja tem mais inadimplências?*
  - *Existe alguma relação entre idade e inadimplência? (**cliente_idade**)*
  - *É possível prever inadimplência através dos dados idade, cidade e produto?*
- **Gere novos dados ou transforme!**

Como iniciar a análise de dados??

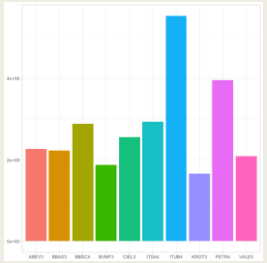


Como iniciar a análise de dados??

- Quais lojas mais vendem?
- Quais produtos mais vendem?
- Quais lojas geram maior receita?
- Quais produtos geram maior receita?
- Existe algum cliente que gera maior receita?
- Qual o tempo médio entre compra e pagamento?
- Existe alguma loja em que esse tempo é menor? E produto?
- Existe alguma receita usando combinação entre produto e loja que mais se destaca?
- Qual produto gera maior inadimplência?
- Qual loja tem maior inadimplência?

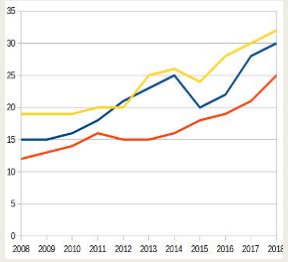
- Existe sazonalidade? por loja? por produto?
- As vendas estão crescendo a cada ano?
- A loja que mais vende é a que mais gera inadimplência?
- Existe alguma relação entre idade e inadimplência?
- É possível prever inadimplência através dos dados idade, cidade e produto?

Gráficos



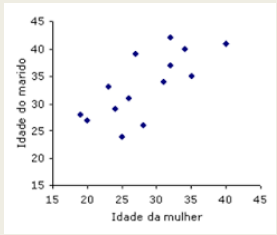
Barras

Usado para categoria ou tempo contra dado numérico
Bom para mostrar comportamentos de anomalias



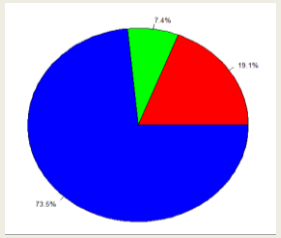
Linhas

Usado para tempo contra dado numérico
Ou análise de 2 dados numéricos
Bom para mostrar tendência ou ruídos



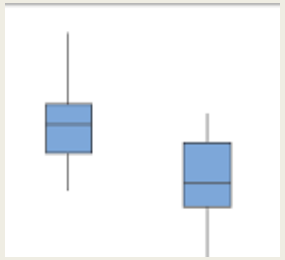
Dispersão

Normalmente usado para análise de 2 dados numéricos (ou tempo em um eixo)
Bom para mostrar correlação entre dados



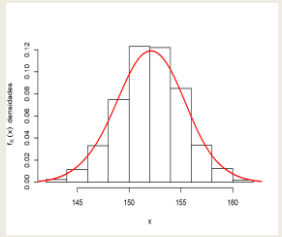
Pizza

Usado para comparar porcentagens de um dado categórico.
Somente usar com porcentagem!



Boxplot

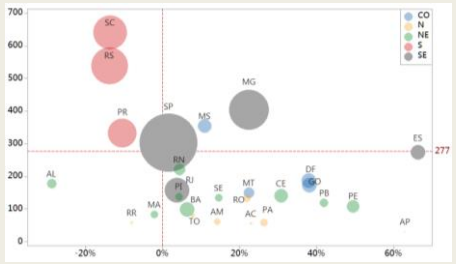
Usado para categoria contra dado numérico
Mostra a variação do dados e outliers
Bom para buscar outliers e avaliar se uma variável impacta no comportamento de outra



Histograma e Curva de Distribuição Normal

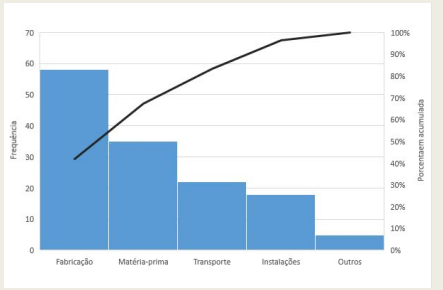
Normalmente usado para análise de 2 dados numéricos
Um dos dados deve ter uma média ou valor esperado

Gráficos



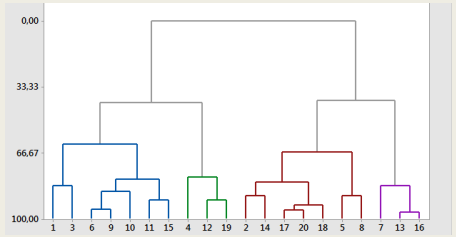
Bolhas

Similar ao gráfico de dispersão, porém uma terceira dimensão é adicionada.



Pareto

Usado para categoria contra dado numérico. Ótimo para identificar se existem categorias que são muito significativas



Dendograma

Mostra a frequência de um dado dentro de diversas categorias

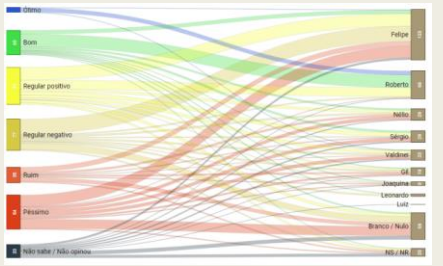
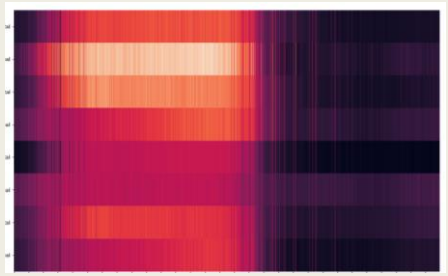


Diagrama de Sankey

Mostra o relacionamento entre 2 variáveis

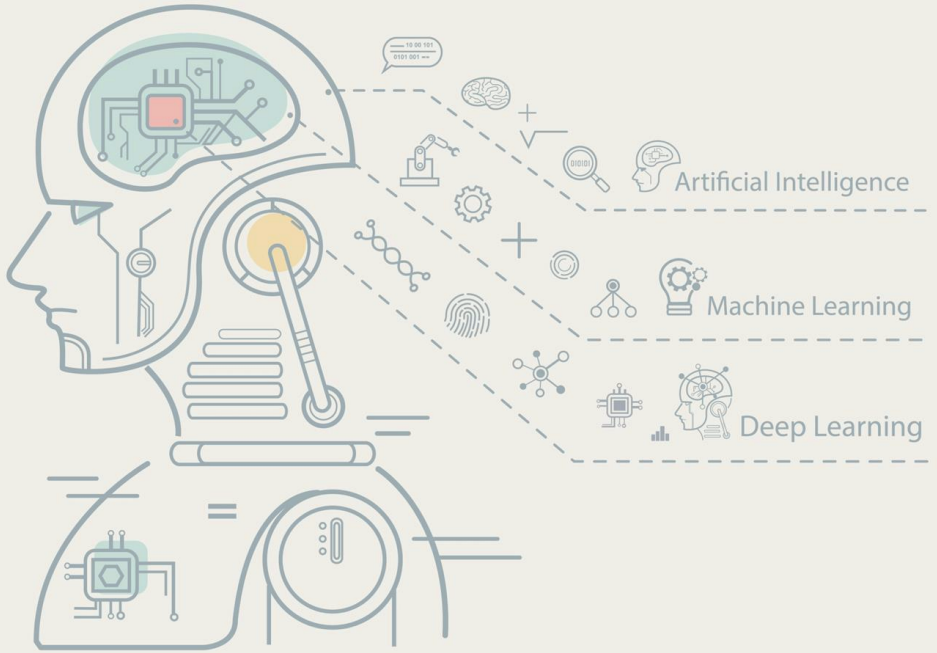


Mapa de calor

Mostra o comportamento de uma variável comparada a outras 2.

Conceitos de ML

- Definição de um alvo
- Normalizar dados
- Dados categóricos
- Dados de treinamento e de teste
- Escolher um modelo
- Matriz de confusão



		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

		Precisão	
		0	**1**
Recall	**0**	VP	FN
	1	FP	VN

Conceitos de ML

Dado 1	Dado 2	Dado 3	Target
A	D	G	0
B	E	H	1
C	F	I	1

model = modelo(parâmetros)

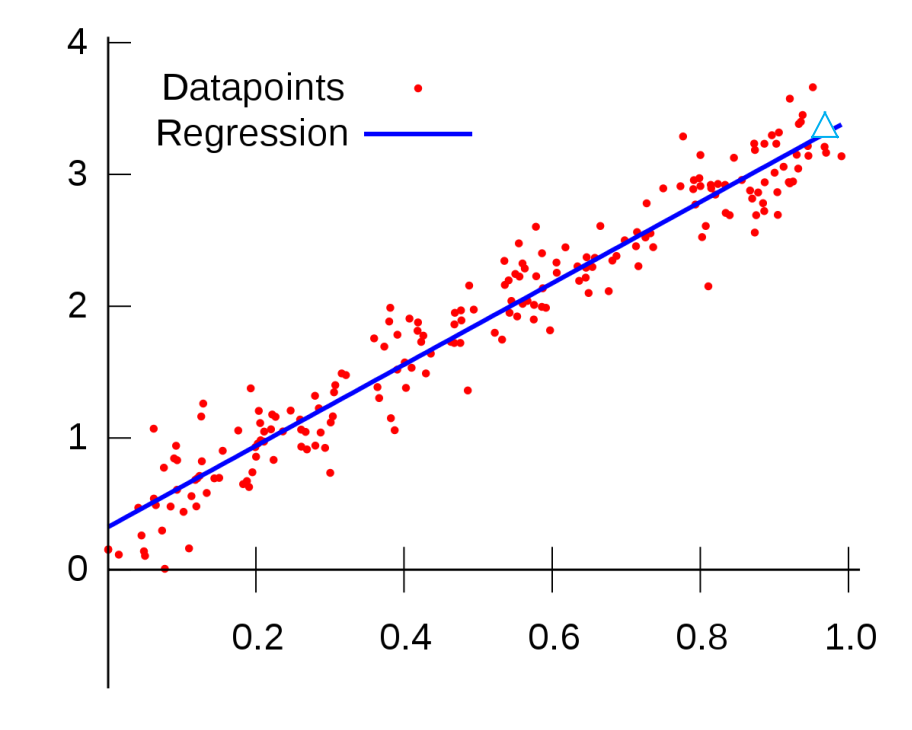
model.fit(X, y)

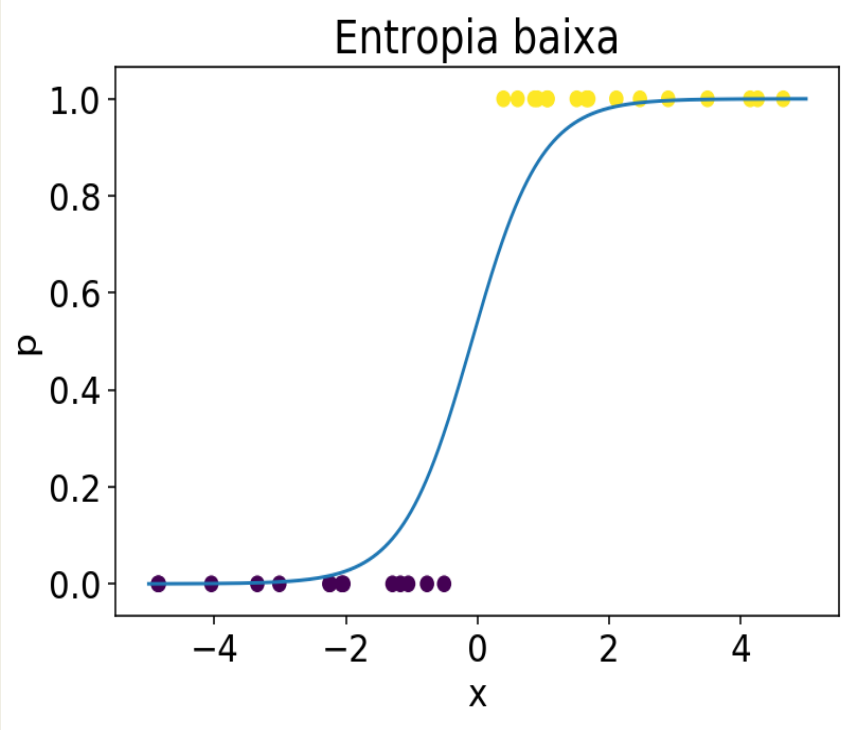
Dado 1	Dado 2	Dado 3	Target
A	E	I	???

model.predict(X)

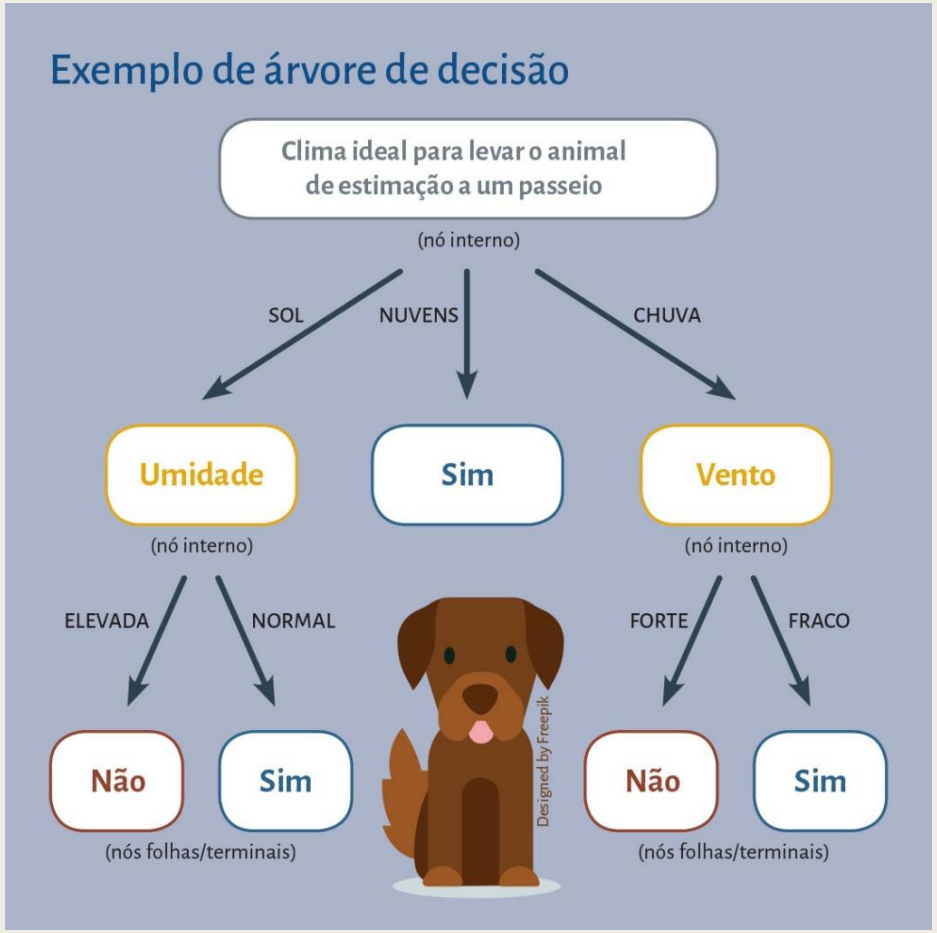
Target = 1

Regressão Linear e Regressão Logística

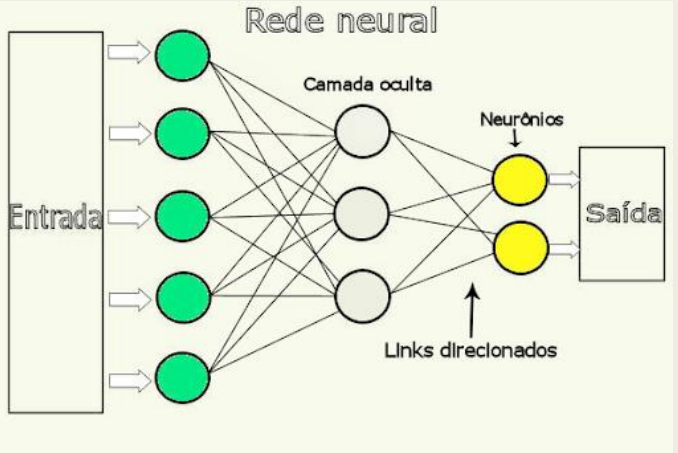


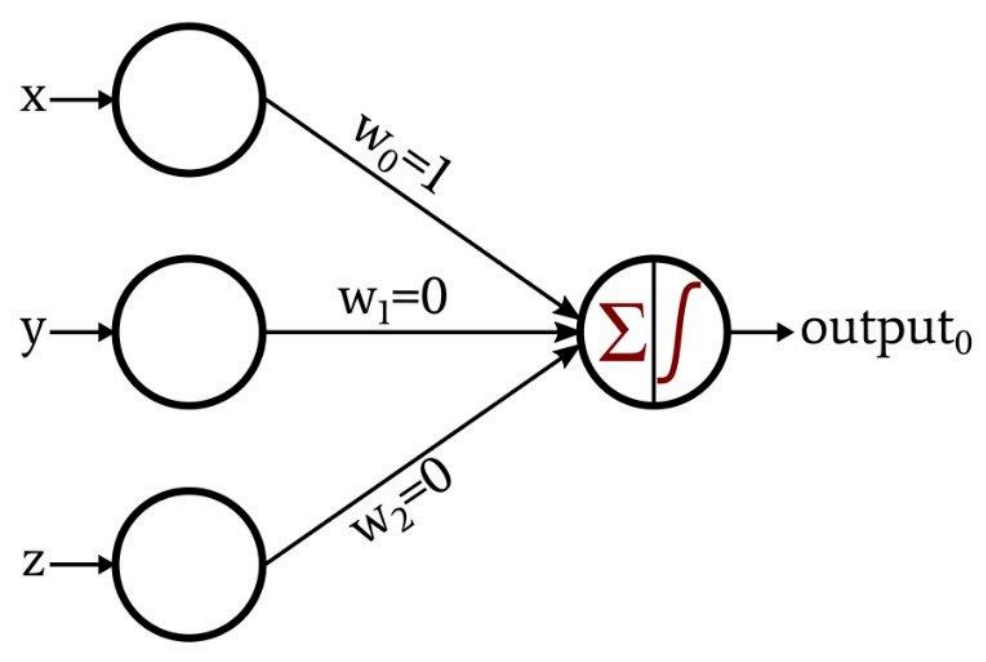


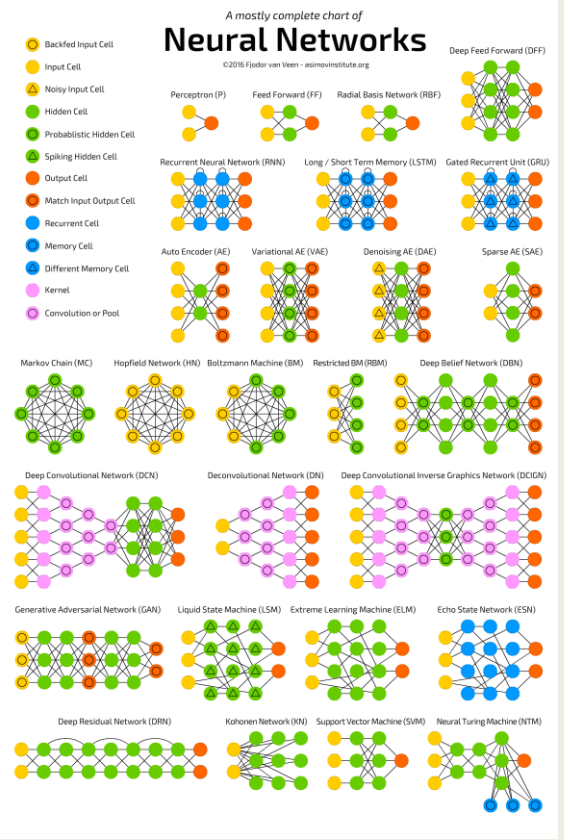
Árvore de decisão



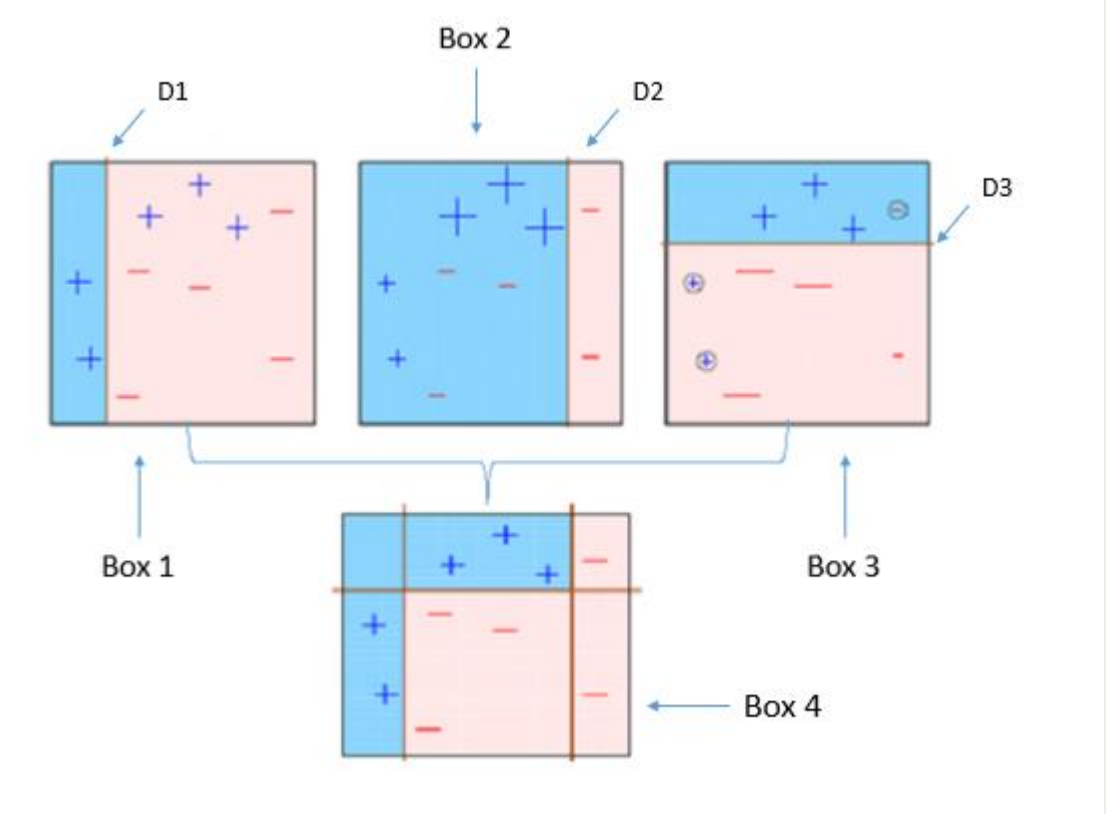
Redes Neurais



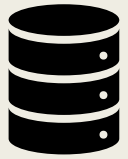




XGBoost



Framework de Trabalho (resumo)



Um problema e dados

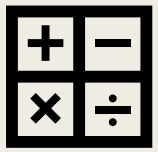
Análises

- Entender o contexto do problema
- Extrair os dados
- Data Cleaning
- Data Fusion
- Feature Engineering
- Gere Perguntas
- Interagir e Visualizar

Ache informações e tire ideias

Machine Learning

- Crie sua base com dados os relevantes
- Defina ou crie uma coluna alvo (target)
- Normalize os dados
- Trate os dados categóricos
- Separe base treino e teste
- Teste modelos de ML!



Gere métricas

Gere dinheiro ou salve vidas!