

SHIFT
FIAP

Ana Raquel



Carreira

- Tecnólogo em banco de dados pela faculdade FIAP.
- MBA em inteligência artificial pela FIAP.
- Mais de 8 anos de experiência como profissional na área de dados tendo atuado em diversos projetos de Banco de Dados, BI, Analytics e Data Science.
- Cientista de dados na FIAP e professora de Machine Learning , Deep Learning, Processamento de Linguagem Natural e Data Viz na FIAP.

Classificação de texto

Classificação de texto

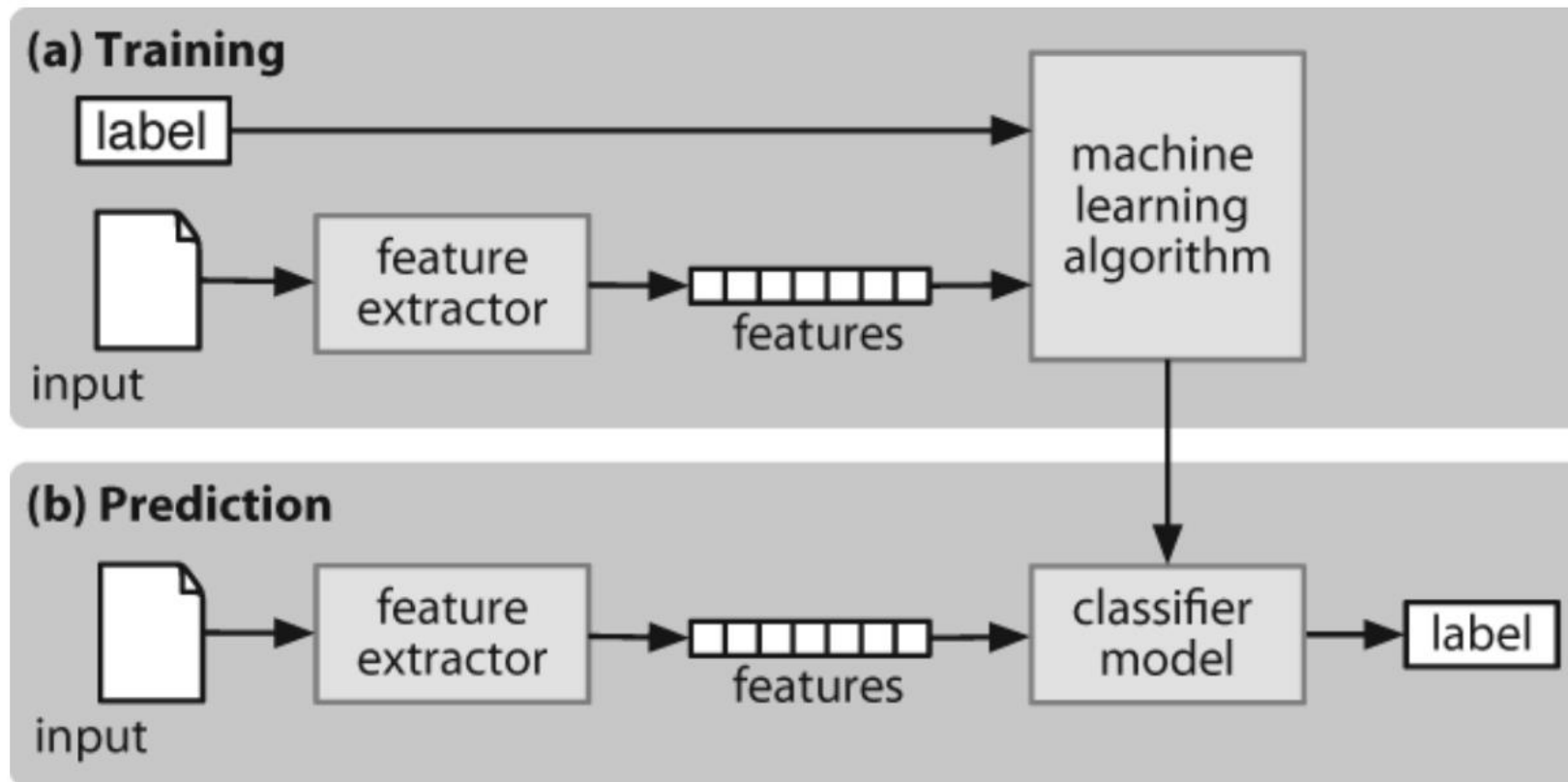
Você já parou para pensar como conseguimos classificar com base em texto?

A classificação é a tarefa de escolher o **rótulo de classe** correto para uma determinada entrada. No básico tarefas de classificação, cada entrada é considerada isoladamente de todas as outras entradas, e o conjunto de rótulos é definido antecipadamente. Alguns exemplos de tarefas de classificação são:

- Decidir se um e-mail é spam ou não.
- Decidir qual é o tópico de um artigo de notícias, a partir de uma lista fixa de áreas temáticas, como “esportes”, “tecnologia” e “política”.
- Classificação de texto positivo e negativo (análise de sentimento).
- Chatbot...



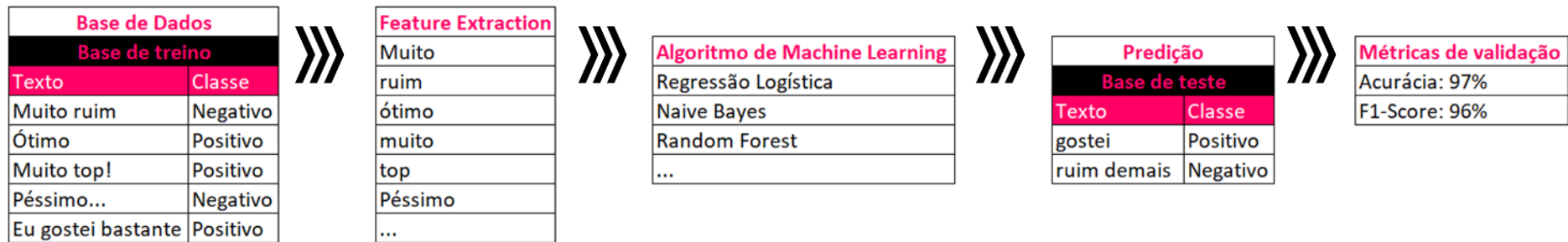
Classificação de texto



Classificação de texto

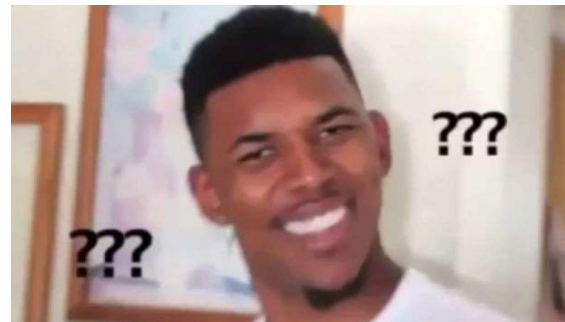
Classificação supervisionada.

Durante o treinamento, um extrator de características é usado para **converter cada valor de entrada para um conjunto de recursos**. Pares de conjuntos de recursos e rótulos são alimentados no algoritmo de aprendizado de máquina para gerar um modelo. Durante a previsão, o mesmo recurso extrator é usado para converter entradas não vistas em conjuntos de recursos. Esses conjuntos de recursos são então inseridos no modelo, que gera rótulos previstos.



Classificação de texto

Para realizar a classificação de texto, é preciso **transformar** o texto em números...mas como eu consigo realizar essa técnica?



Base de Dados	
Base de treino	
Texto	Classe
Muito ruim	Negativo
Ótimo	Positivo
Muito top!	Positivo
Péssimo...	Negativo
Eu gostei bastante	Positivo



Feature Extraction
Muito
ruim
ótimo
muito
top
Péssimo
..



Algoritmo de Machine Learning
Regressão Logística
Naive Bayes
Random Forest
...



Predição	
Base de teste	
Texto	Classe
gostei	Positivo
ruim demais	Negativo



Métricas de validação
Acurácia: 97%
F1-Score: 96%

Bag of Words (BOW)

O Tipo de representação Bag of Words utiliza um **vetor de contagem de palavras** para representar um documento. Esse modelo apenas realiza a contagem de cada palavra e não a ordem em que cada uma aparece no documento. Com essa técnica, **vale ressaltar que o contexto e rases são ignorados**.

Eu gosto muito de café coado, mas eu também gosto muito de café com leite. Na verdade, tudo que tem café eu gosto!



Café	2
gosto	2
muito	2
eu	2
também	1
que	1
mas	1
leite	1
na	1
verdade	1
tem	1
coado	1
com	1

TF-IDF

Agora vamos pensar no seguinte questionamento:

Todas as palavras de um documento são igualmente importantes?

A técnica de TF-IDF (Term Frequency – Inverse Document Frequency) utiliza um conceito de **frequência de termo que calcula a proporção de um termo em um documento em relação ao número total de termos** nesse documento. Sendo assim, palavras muito frequentes começa a dominar o documento. Um ponto interessante nessa técnica é que o IDF (inverse document frequency) **calcula a importância de um termo**, classificando assim palavras frequentes vs palavras raras.

Palavras comuns recebem 0 e palavras raras recebem 1.

$$TF-IDF = TF(t, d) \times IDF(t)$$

Documento analisado

Frequência do termo

Inverse document frequency
(Frequência de documento inversa)

$\text{Log} \frac{1 + n}{1 + df(d, t)} + 1$

•

●

+

•

 $+$

X

✕

■ ■ ■

Normalização de texto

Você percebeu que no nosso exemplo anterior utilizamos as stop words para deixar o texto mais “limpo” e não calcular frequência de palavras irrelevantes no documento? Sem contar que a pontuação também não apareceu em nossa lista de frequência.

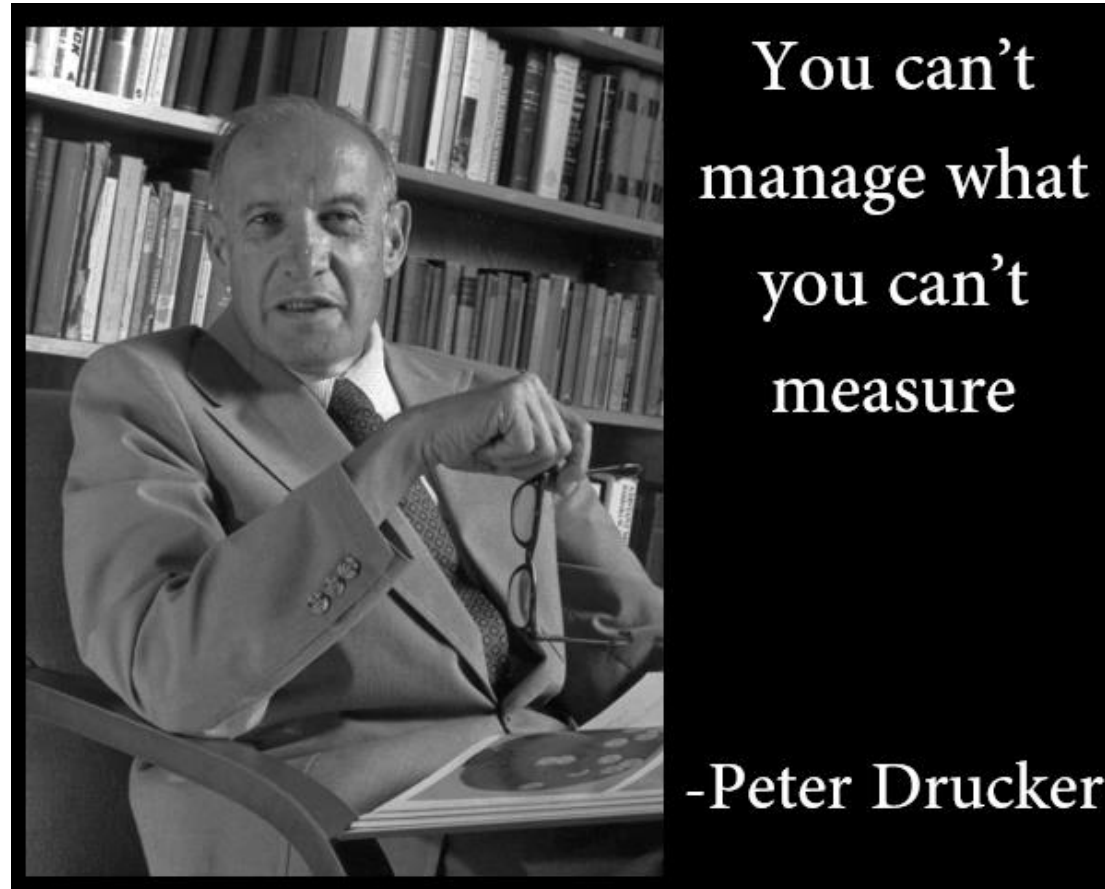
Utilizar técnicas de normalização de texto antes de construir um modelo é obrigatório para obter sucesso nos resultados, temos abaixo um pipeline de transformações a serem realizadas antes de contabilizar a frequência das palavras;

- Tokenização
- Stop words
- Remoção de acentuação e pontuação
- Lematização ou Stematização
- Lowercasing

Métricas de **avaliação**

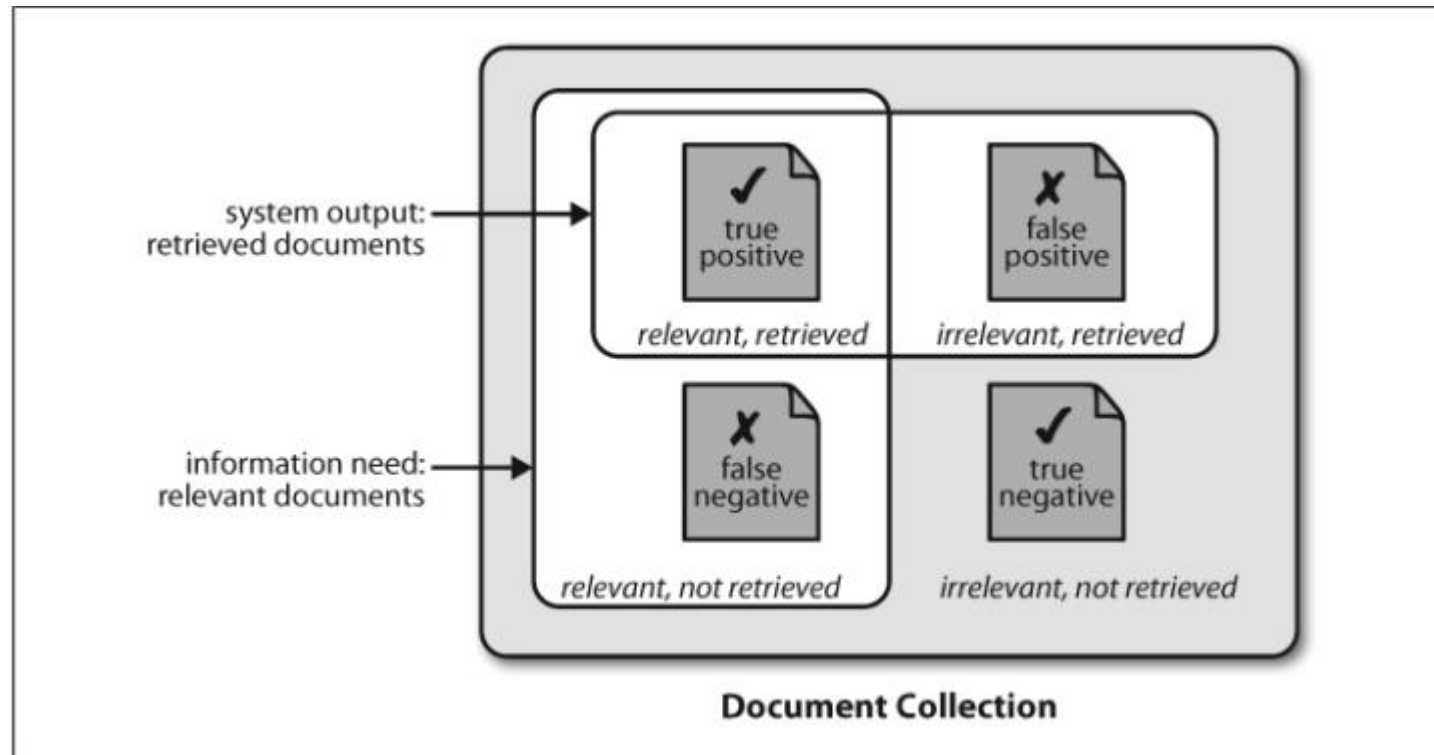
Métricas de avaliação

Mas por que precisamos validar o modelo?



Métricas de avaliação

Temos como as mais comuns métricas de avaliação de modelos: **Acurácia**, **Precision**, **Recall** e **F1 Score**. Elas são obtidas a partir da **Matriz de Confusão**, apresentada abaixo:



Acurácia

Mede a porcentagem de entradas no conjunto de teste que o classificador rotulou corretamente.

Por exemplo, um classificador de gênero de nome que prevê o nome correto 60 vezes em um conjunto de teste contendo 80 nomes teriam uma precisão de $60/80 = 75\%$.

Ao interpretar a pontuação de precisão de um classificador, é importante considerar a frequências dos rótulos de classes individuais no conjunto de teste.



Precise but
not accurate



Accurate
and precise



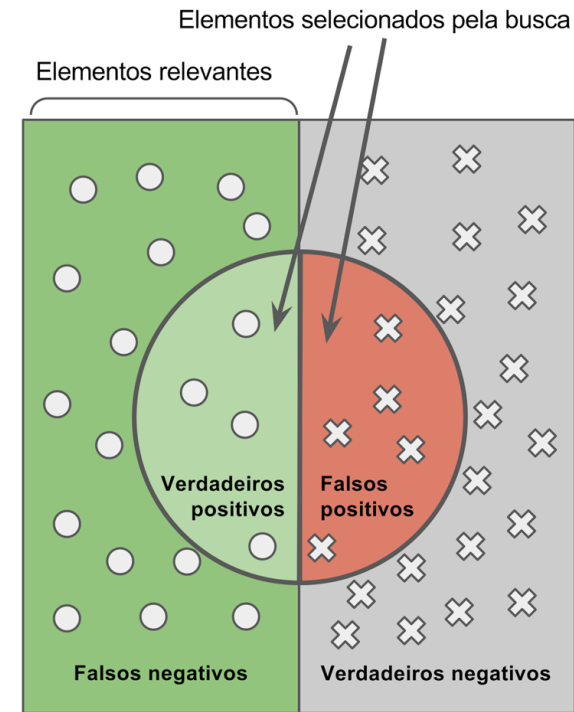
Accurate but
not precise





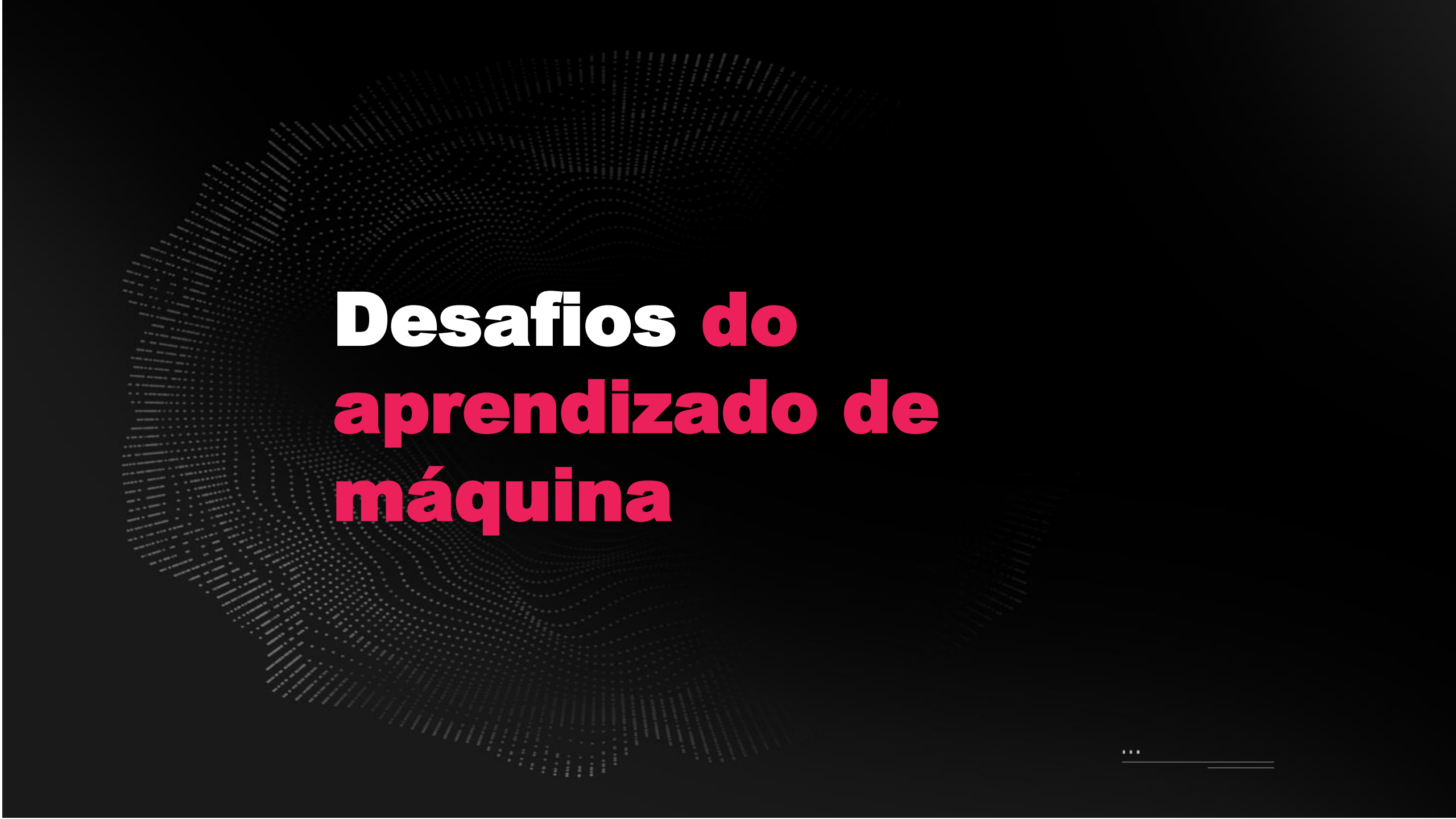
Not accurate
or precise

Precision e Recall

- **Precisão**, que indica quantos dos itens que identificamos foram relevantes, é $TP/(TP+FP)$.
- **Recall**, que indica quantos dos itens relevantes que identificamos, é $TP/(TP+FN)$.
- **F-Score**, que combina a precisão e a recuperação para fornecer uma pontuação única, é definida como **a média harmônica da precisão e revocação**.



Precisão = 	Revocação = 
"Quantos elementos selecionados são relevantes?"	"Quantos elementos relevantes foram selecionados?"



Desafios do aprendizado de máquina

Desafios do aprendizado de máquina

Mesmo aplicando as técnicas de limpeza na base, podemos encontrar um resultado não tão satisfatório ao desenvolver o modelo...você já parou para se questionar o que pode ter acontecido?

Podemos ter alguns impedimentos que podem ocasionar resultados ruins em nossos algoritmos, sendo eles:

- **Algoritmos ruins.**
- **Dados ruins**



me cleaning
the data

me building
a model

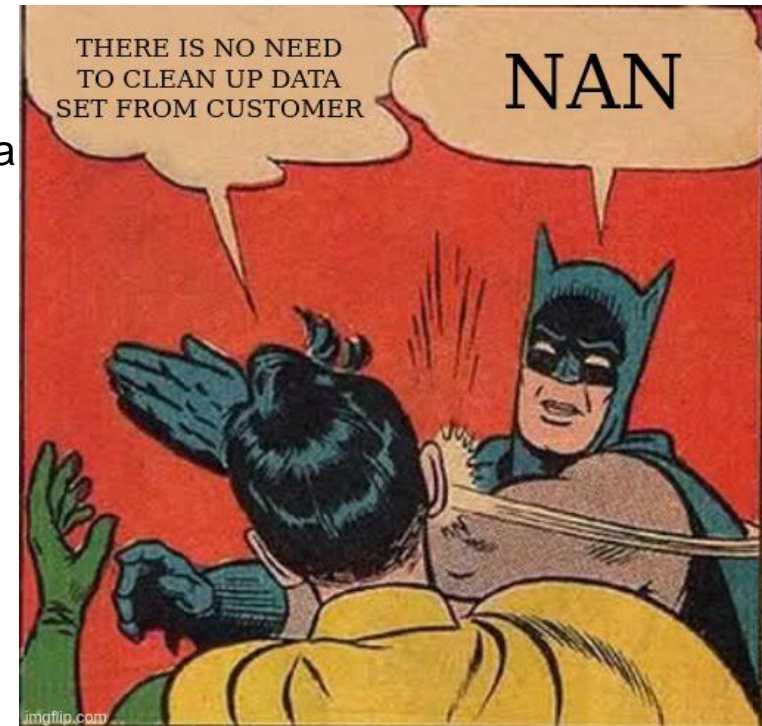
Dados ruins

Quantidade insuficiente de dados

Com uma **amostra de dados muito pequena**, existirá um "**ruído de amostragem**" e se houver uma **amostra muito grande com dados não representativos**, o método de amostragem também pode ser falho (**Viés de amostragem**).

Um dos grandes desafios ao construir um modelo de machine learning, é a **preparação da base de dados**.

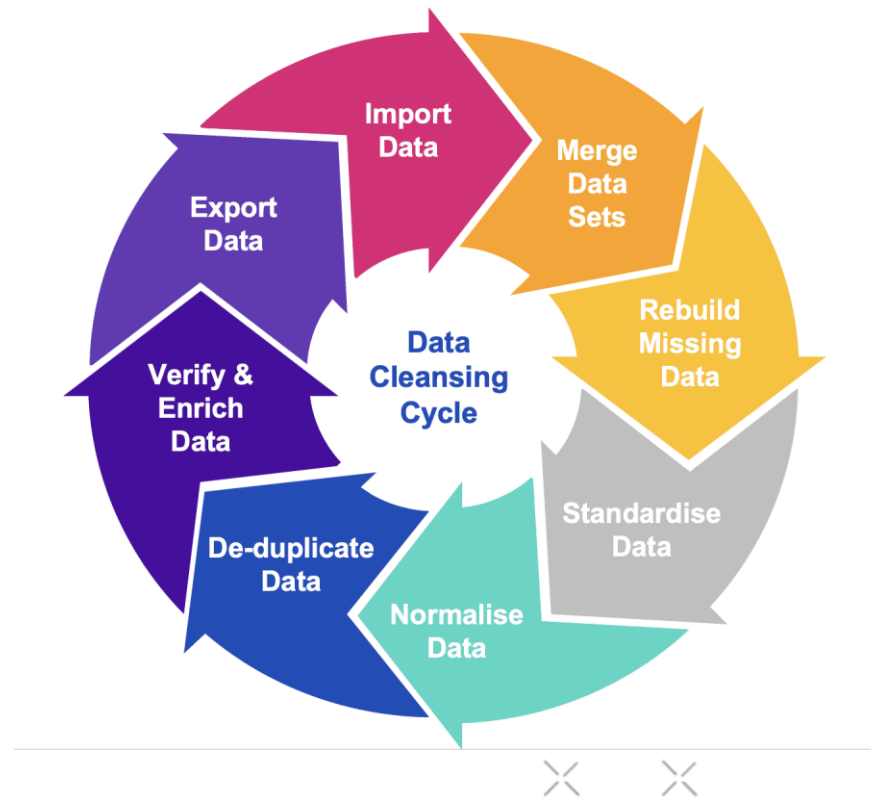
No mundo real, infelizmente não temos as “bases do kaggle” que muitas delas possuem dados estruturados e já no formato correto para aplicar no modelo de machine learning.



Dados ruins

Dados de baixa qualidade

Aqui é preciso se dedicar a limpeza dos dados, uma base não consistente pode impactar na detecção de padrões.

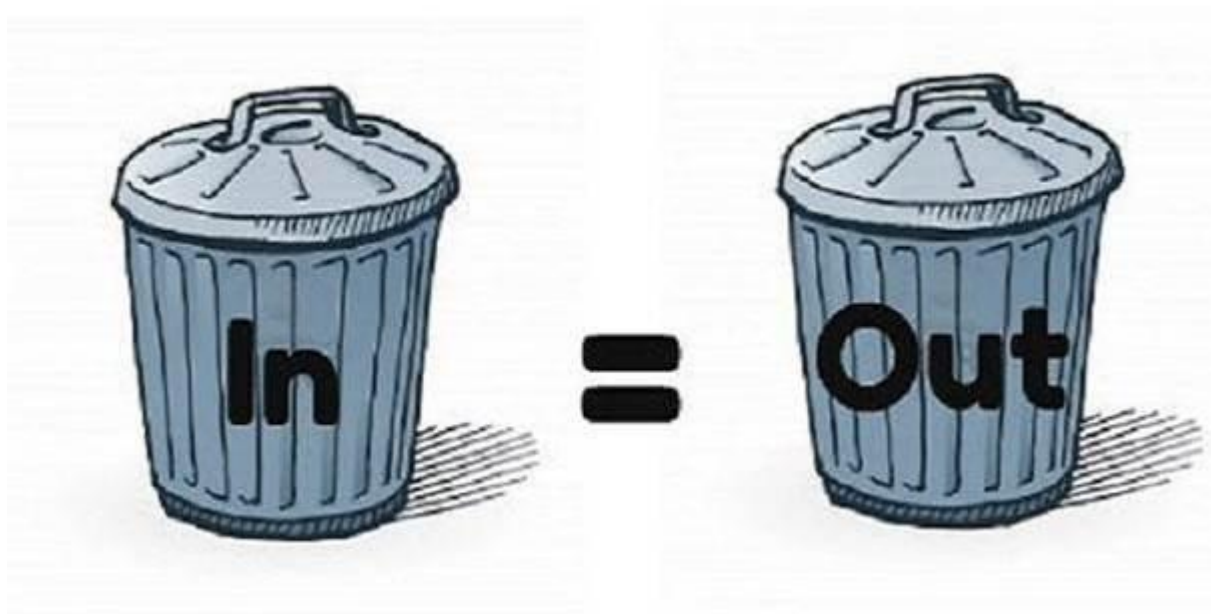


Dados ruins

Características irrelevantes

Entra lixo, sai lixo!

Atenção aos dados que entram no seu modelo! A dica aqui é a dedicação na etapa **de feature engineering** ou **técnicas de redução de dimensionalidade**.



Algoritmos ruins

Sobreajuste nos dados (overfitting)

Quando o seu modelo funciona muito bem com os dados de treinamento **mas não generaliza bem novos dados de entrada**. Isso pode acontecer quando o modelo é muito complexo em relação ao ruído e quantidade.

Como **solução** podemos pensar aqui em:

- *Simplificar o modelo.*
- *Coletar mais dados.*
- *Reduzir o ruído* (exemplo, remover outliers).
- *Regularização*: Chamamos de regularização quando restringimos um modelo para simplificar e reduzir o risco de reajuste dos dados. A regularização pode ajudar a generalizar melhor o modelo em novos exemplos de dados.



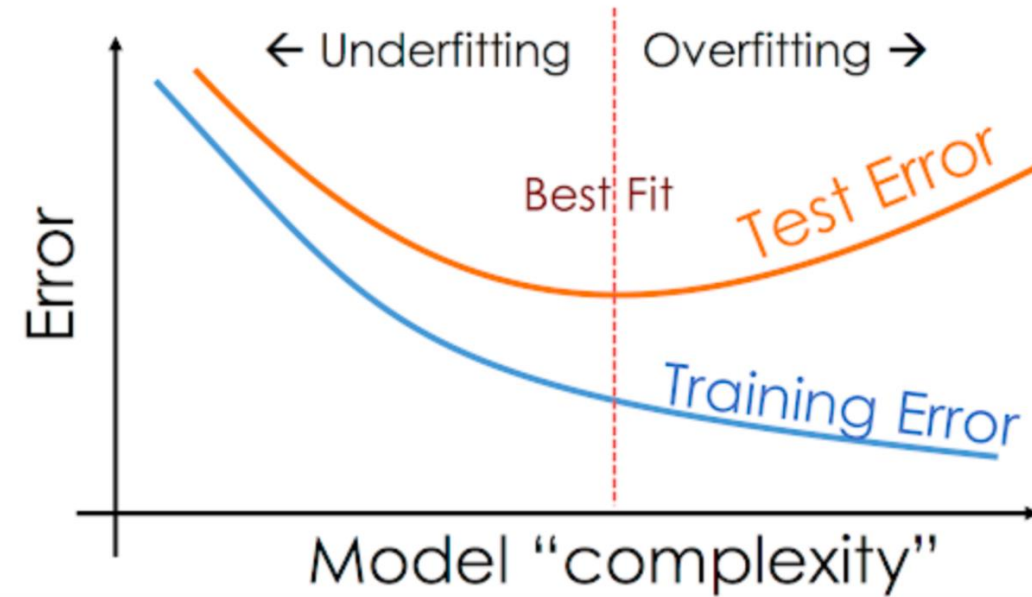
Algoritmos ruins

Subajuste dos dados (underfitting)

Nesse caso seu **modelo ficou muito simples** ao ponto de não aprender corretamente os dados.

Algumas **soluções**:

- *Selecionar um modelo mais poderoso.*
- *Feature engineering.*
- *Reduzir as regularizações.*

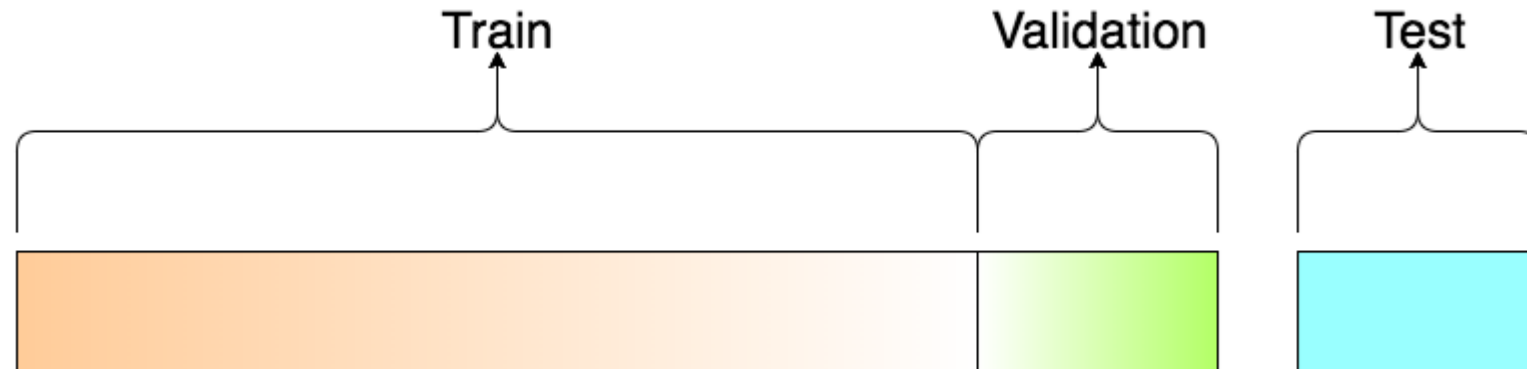


Algoritmos ruins

Falta de teste e validação do algoritmo

Base de treino, teste e validação cruzada

Para medir o erro da generalização, utilize a técnica de treinar vários hiperparâmetros utilizando o conjunto de treinamento, selecione os hiperparâmetros que melhor se adapta no conjunto de validação.



Obrigada!

Ana Raquel



[linkedin.com/ana-raquel-fernandes-cunha](https://www.linkedin.com/ana-raquel-fernandes-cunha)

Copyright © 2023 | Ana Raquel Fernandes Cunha

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento é expressamente proibido sem consentimento formal, por escrito, do professor/autor.