Disciplina: Machine Learning

Aula 02: Conceito de Machine Learning

Profª. Wanderlan Carvalho de Albuquerque

Agenda

- O que é *Machine Learning*?
- Conceitos correlacionados

O que é *Machine Learning*?

O que é o Machine Learning?

- É um método de **análise de dados** que automatiza a construção de **modelos analíticos**. É um ramo da **inteligência artificial** baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana.
- Diferente do ***Deep Learning*** **(ou DL)**, o ML necessita de um treinamento prévio feito de formas de aprendizado supervisionadas, não supervisionadas, semi-supervisionado, etc.
- DL é uma especialização do ML e necessita de grande poder computacional para poder ser viável*.
- Tipos de dados podem ser: **imagens, sons e textos**.

*depende da necessidade do usuário e complexidade do dataset

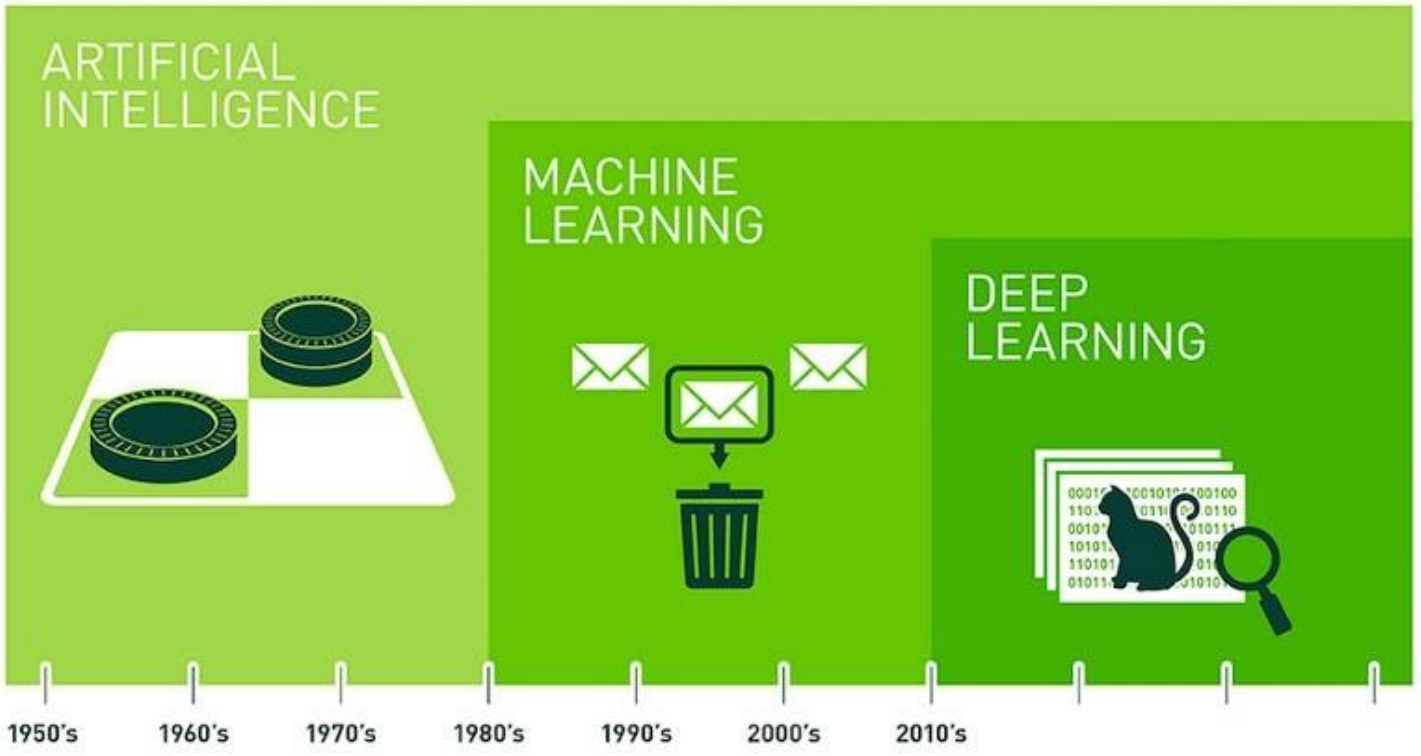
Por que o ML só surgiu agora?

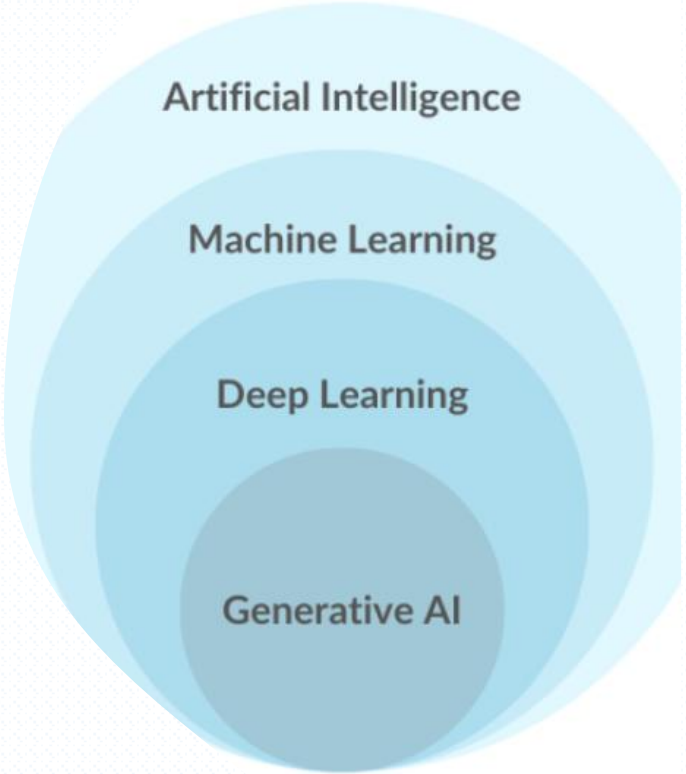
- É um erro comum dizer isso...
- Algoritmos de ML já remotam desde 1960 (e até antes). Na década de 90, os correios americanos já usavam ML para detectar números de CEP e escrita em correspondências.
- O ML veio a se tornar bem famoso (somente agora) por 3 motivos:
  - Os métodos de ML são mais precisos que os próprios humanos, desde 2014 em diante;
  - Processamento de GPU's permitem criar classificadores em um tempo viável;
  - Grande quantidade da dados classificados (*Big Data*) só começaram a serem disponibilizados com a globalização da Internet e o surgimento de redes mais rápidas e eficientes.

Aonde é utilizado

- Medicina
- Computação afetiva
- Mecanismos de busca
- Blockchain e finanças
- Moda
- Detector de fraudes
- Posicionamento futuro (predizer em que posição você irá se mover)
- Industria
- Visão computacional (para reconhecimento de objetos)
- Mineração de dados
- Marketing

O que é o Machine Learning?





https://medium.com/data-science-brigade/a-diferen%C3%A7a-entre-intelig%C3%AAncia-artificial-machine-learning-e-deep-learning-930b5cc2aa42

ACADEMY TALENTOS

Conceitos Correlacionados

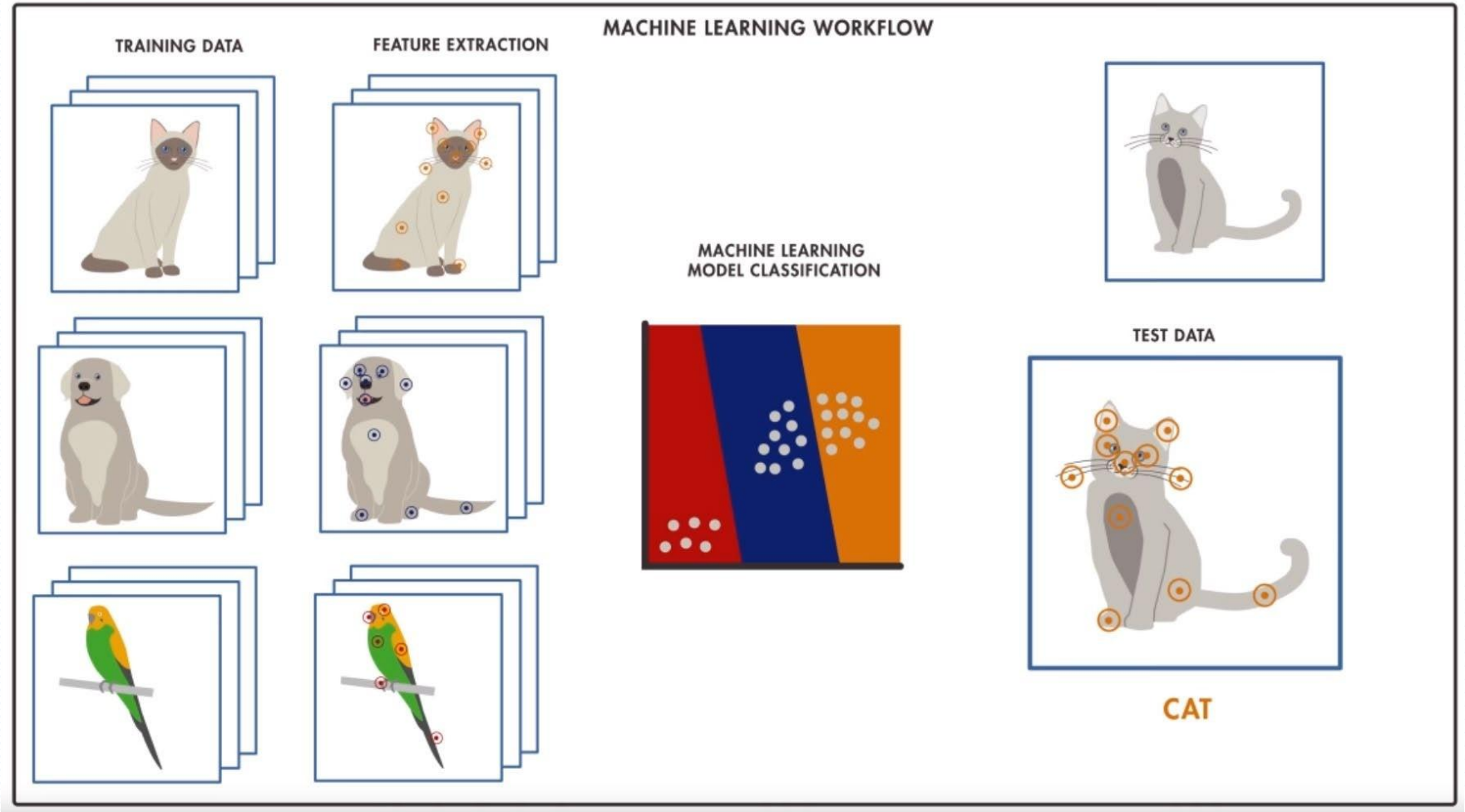
Comparações

Algoritmo Computacional	*Machine Learning*
Baseado em entradas: dados atuais	Baseado em dados históricos
Utiliza apenas algoritmos	Algoritmo + Modelo
100% de performance	Não podemos esperar 100%
Performance constante	Performance pode variar
Atende qualquer negócio	Focado a um negócio específico
Não precisa aprender	Precisa aprender e reaprender

Comparações

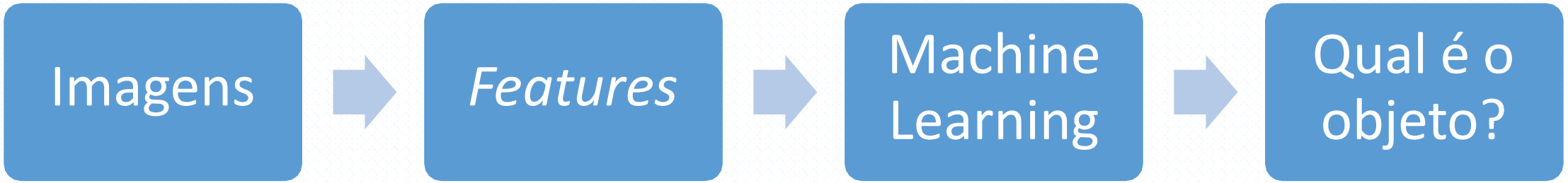
	Machine Learning	*Deep Learning*
Dataset (DB) de treino	Pequeno	Grande
Escolha de suas próprias características	Sim	Não
No. de classificadores disponíveis	Muitos	Poucos
Tempo de treinamento	Curto	Longo
Necessita de hardware poderoso	Não	Sim

Exemplo: Gatos versus Cachorros



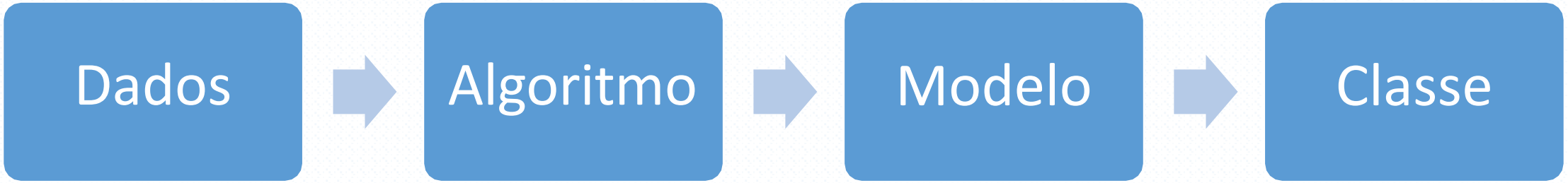
Introduction to Deep Learning: Machine Learning vs Deep Learning
https://www.youtube.com/watch?v=-SgkLEuhfbg

Fluxo de Operação

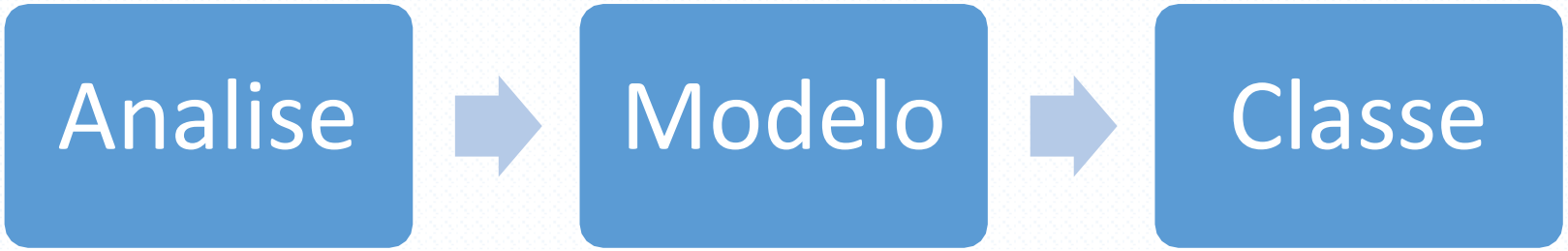


Modelo

Primeira rodada para criar o modelo



Após o modelo ser criado



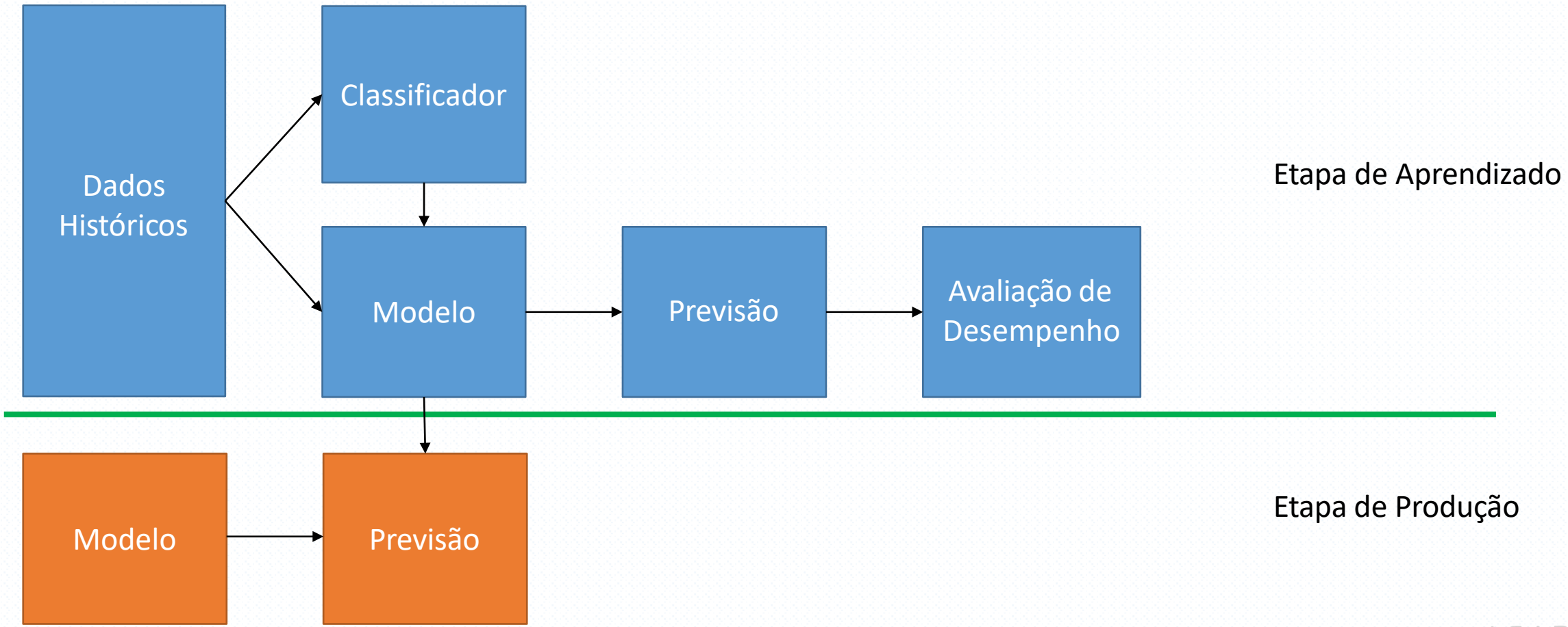
Medindo o Desempenho do Modelo

- Divide-se o *dataset* em duas partes:
  - Treino: parte que o algoritmo processa e cria o modelo.
  - Teste: dados que são submetidos ao modelo e verifica-se a precisão.

- Essa separação pode ser feita de quatro formas:
  - Mesmo *Dataset*;
  - *Hold-Out*;
  - Sub-amostragem Aleatória; e
  - Validação Cruzada.

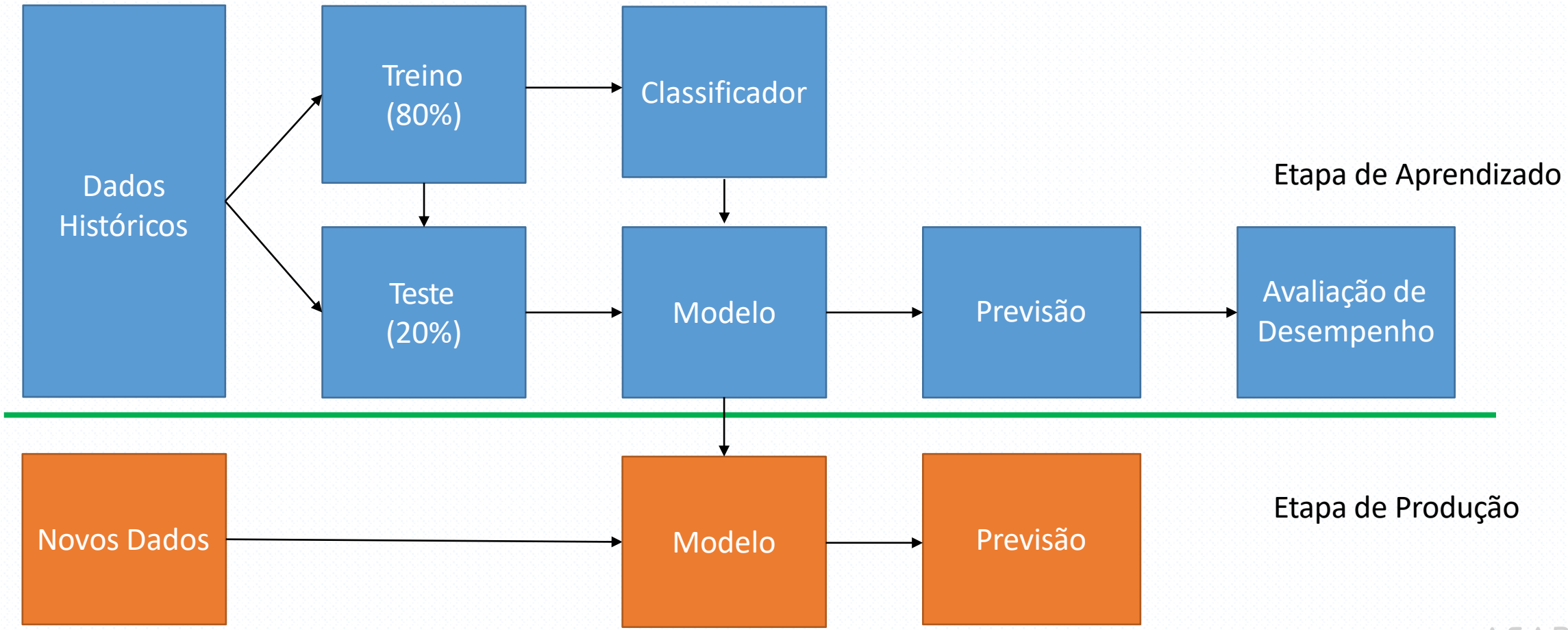
Medindo o Desempenho do Modelo

- Mesmo *Dataset*



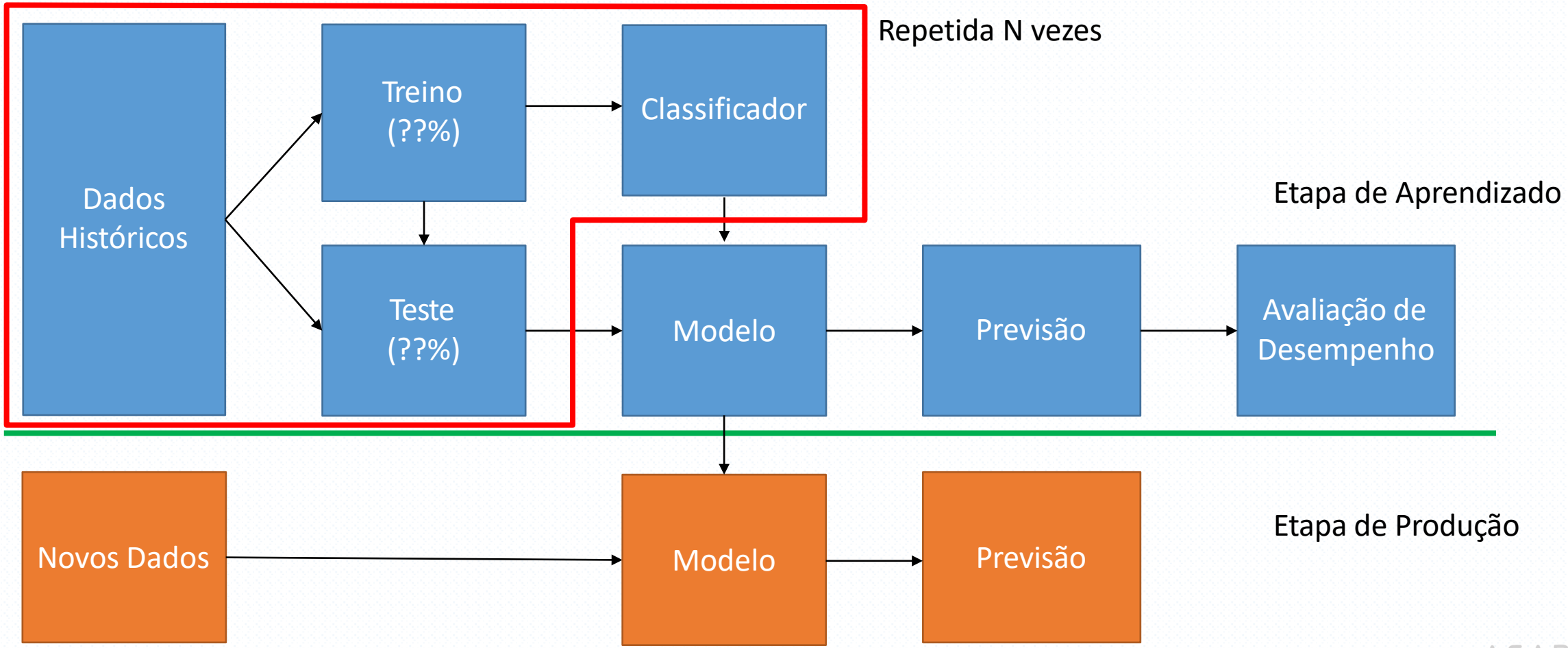
Medindo o Desempenho do Modelo

- *Hold-Out*



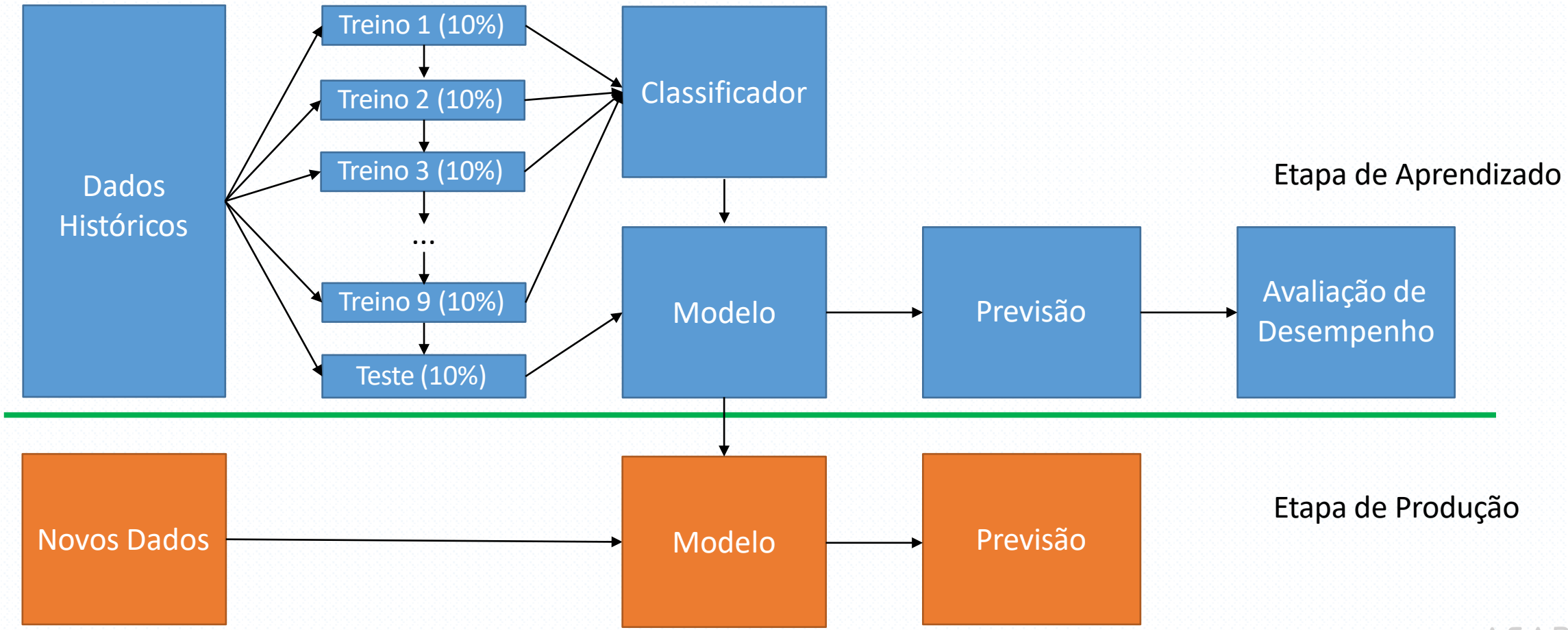
Medindo o Desempenho do Modelo

- Sub-Amostragem Aleatória



Medindo o Desempenho do Modelo

- Validação Cruzada



Matriz de confusão

- Também chamada de **Tabela de Contingência**. Ela é responsável por mostrar os resultados do nosso treino.

		Classe Verdadeira	
		+	-
Predição	+	Verdadeiro Positivo (TP)	Falso Positivo (FP)
	-	Falso Negativo (FN)	Verdadeiro Negativo (TN)

Erro Tipo 1

Erro Tipo 2 (Fatal!!)

Tipos de Erros

- **Erro tipo I**
- O primeiro tipo de erro é a rejeição de uma hipótese nula verdadeira como resultado de um procedimento de teste. Esse tipo de erro é chamado de erro do tipo I e às vezes é chamado de erro do primeiro tipo.
- Em termos do exemplo de um tribunal, um erro do tipo I corresponde à condenação de um réu inocente.

Tipos de Erros

- **Erro tipo II**
- O segundo tipo de erro é a falha em rejeitar uma hipótese nula falsa como resultado de um procedimento de teste. Esse tipo de erro é denominado erro do tipo II e também denominado erro do segundo tipo.
- Em termos do exemplo de um tribunal, um erro do tipo II corresponde à absolvição de um criminoso.

Tipos de Erros

- **Falso positivo e falso negativo**
- Em termos de falsos positivos e falsos negativos, um resultado positivo corresponde à rejeição da hipótese nula, enquanto um resultado negativo corresponde à não rejeição da hipótese nula; "falso" significa que a conclusão tirada está incorreta.
- Assim, um erro do tipo I é equivalente a um falso positivo e um erro do tipo II é equivalente a um falso negativo.

Tipos de Erros

Para Matriz da Confusão !

- **Falso Positivo (False Positive)**: Ocorre quando o modelo prediz que uma condição ou evento está presente, mas na realidade, não está. Por exemplo, se um teste de diagnóstico prediz que uma pessoa tem uma doença quando, na verdade, ela não tem, isso é um falso positivo.
- **Falso Negativo (False Negative)**: Ocorre quando o modelo prediz que uma condição ou evento não está presente, mas na realidade, está. Por exemplo, se um teste de diagnóstico prediz que uma pessoa não tem uma doença quando, na verdade, ela tem, isso é um falso negativo.

Tipos de Erros

Tabela de tipos de erros

Relações tabularizadas entre verdade / falsidade da hipótese nula e resultados do teste:

Tabela de tipos de erros		**Hipótese nula (H_0) é**	
		Verdade	**Falso**
Decisão sobre hipótese nula (H_0)	**Não rejeite**	Inferência correta (verdadeiro negativo) (probabilidade = $1-\alpha$)	Erro tipo II (falso negativo) (probabilidade = β)
	Rejeitar	Erro tipo I (falso positivo) (probabilidade = α)	Inferência correta (verdadeiro positivo) (probabilidade = $1-\beta$)

https://en.wikipedia.org/wiki/Type_I_and_type_II_errors#Type_I_error

Matriz de confusão

- **Exemplo**: temos 13 imagens com cães e gatos. Gatos pertencem a classe 1 e cachorros na 0. A classe verdadeira é: [1,1,1,1,1,1,1,1,0,0,0,0,0].
- O classificador prediz dessa forma: [0,0,0,1,1,1,1,1,0,0,0,1,1]

		Classe Verdadeira	
		Gatos	**Cães**
Predição	**Gatos**	**5**	**2**
	Cães	**3**	**3**

Matriz de confusão - Métricas

		Condição verdadeira				
	População total	Condição positiva	Condição negativa	Prevalência $= \frac{\Sigma \text{ Condição positiva}}{\Sigma \text{ População total}}$	Precisão (ACC) $= \frac{\Sigma \text{ Verdadeiro positivo} + \Sigma \text{ Verdadeiro negativo}}{\Sigma \text{ População total}}$	
Condição prevista	Condição prevista positiva	**Verdadeiro positivo**	**Falso positivo**, erro Tipo I	Valor preditivo positivo (PPV), precisão $= \frac{\Sigma \text{ Verdadeiro positivo}}{\Sigma \text{ Condição prevista positiva}}$	Taxa de descoberta falsa (FDR) $= \frac{\Sigma \text{ Falso positivo}}{\Sigma \text{ Condição prevista positiva}}$	
	Condição prevista negativa	**Falso negativo**, erro Tipo II	**Verdadeiro negativo**	Taxa de falsa omissão (FOR) $= \frac{\Sigma \text{ Falso negativo}}{\Sigma \text{ Condição prevista negativa}}$	Valor preditivo negativo (NPV) $= \frac{\Sigma \text{ Verdadeiro negativo}}{\Sigma \text{ Condição prevista negativa}}$	
		Taxa positiva verdadeira (TPR), Recall, Sensibilidade, probabilidade de detecção, Potência $= \frac{\Sigma \text{ Verdadeiro positivo}}{\Sigma \text{ Condição positiva}}$	Taxa de falsos positivos (FPR), Fall-out, probabilidade de falso alarme $= \frac{\Sigma \text{ Falso positivo}}{\Sigma \text{ Condição negativa}}$	Razão de verossimilhança positiva (LR +) $= \frac{TPR}{FPR}$	Odds ratio de diagnóstico (DOR) $= \frac{LR+}{LR-}$	Pontuação $F_1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Rechamada}}{\text{Precisão} + \text{recall}}$
		Taxa de falsos negativos (FNR), taxa de falha $= \frac{\Sigma \text{ Falso negativo}}{\Sigma \text{ Condição positiva}}$	Especificidade (SPC), seletividade, taxa negativa verdadeira (TNR) $= \frac{\Sigma \text{ Verdadeiro negativo}}{\Sigma \text{ Condição negativa}}$	Razão de verossimilhança negativa (LR−) $= \frac{FNR}{TNR}$		

https://en.wikipedia.org/wiki/Confusion_matrix

Matriz de confusão - Métricas

Terminologia e derivações de uma matriz de confusão

condição positiva (P)
o número de casos reais positivos nos dados
condição negativa (N)
o número de casos reais negativos nos dados

verdadeiro positivo (TP)
eqv. com acerto
verdadeiro negativo (TN)
eqv. com rejeição correta
falso positivo (FP)
eqv. com falso alarme , erro Tipo I
falso negativo (FN)
eqv. com falha, erro tipo II

sensibilidade , recall , taxa de acerto ou taxa positiva verdadeira (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

especificidade , seletividade ou taxa negativa verdadeira (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = 1 - \mathrm{FPR}$$

precisão ou valor preditivo positivo (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathrm{FDR}$$

valor preditivo negativo (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} = 1 - \mathrm{FOR}$$

taxa de falha ou taxa de falso negativo (FNR)

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}} = 1 - \mathrm{TPR}$$

taxa de queda ou falsos positivos (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{TNR}$$

taxa de descoberta falsa (FDR)

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} = 1 - \mathrm{PPV}$$

taxa de falsa omissão (FOR)

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{NPV}$$

Limiar de prevalência (PT)

$$PT = \frac{\sqrt{TPR(-TNR+1)} + TNR - 1}{(TPR + TNR - 1)}$$

Pontuação de ameaça (TS) ou índice de sucesso crítico (CSI)

$$\mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

precisão (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

precisão balanceada (BA)

$$\mathrm{BA} = \frac{TPR + TNR}{2}$$

Pontuação F1

é a média harmônica de precisão e sensibilidade

$$\mathrm{F_1} = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Coeficiente de correlação de Matthews (MCC)

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

Índice de Fowlkes-Mallows (FM)

$$\mathrm{FM} = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} = \sqrt{PPV \cdot TPR}$$

informação ou informação do bookmaker (BM)

$$\mathrm{BM} = \mathrm{TPR} + \mathrm{TNR} - 1$$

marcação (MK) ou deltaP

$$\mathrm{MK} = \mathrm{PPV} + \mathrm{NPV} - 1$$

Fontes: Fawcett (2006), [1] Powers (2011), [2] Ting (2011), [3] e CAWCR [4] Chicco & Jurman (2020), [5] Tharwat (2018). [6]

https://en.wikipedia.org/wiki/Confusion_matrix

Matriz de confusão – Verificando Precisão

		Classe Verdadeira	
		Gatos	**Cães**
Predição	**Gatos**	**5 (TP)**	**2 (FP)**
	Cães	**3 (FN)**	**3 (TN)**

- Prevalência ou Taxa de Acertos = $\frac{TP+TN}{\sum populatppon} = \frac{(5+3)}{13} = 61{,}5\%$
- Taxa de Erros = $\frac{FP+FN}{\sum populatppon}\% = \frac{(2+3)}{13} = 38{,}5\%$
- Precisão = $\frac{TP}{TP+FP} = \frac{5}{7} = 71{,}43\%$
- Lembrança, sensibilidade ou Positivo Verdadeiros = $\frac{TP}{TP+FN} = \frac{5}{8} = 62{,}5\%$
- Especificidade ou Negativo Verdadeiros = $\frac{TN}{TN+FP} = \frac{3}{5} = 60\%$
- Falso Positivos = $\frac{FP}{FP+TN} = \frac{2}{5} = 40\%$
- Falso Negativos = $\frac{FN}{FN+TP} = \frac{3}{8} = 37{,}5\%$

Matriz de confusão – Sensibilidade

Sensibilidade: também conhecida como recall ou taxa de verdadeiros positivos (TPR - True Positive Rate), é uma métrica que mede a proporção de exemplos positivos que foram corretamente identificados pelo modelo em relação ao total de exemplos positivos reais

Matriz de confusão – Sensibilidade

A fórmula para calcular a sensibilidade é:

$$\text{Sensibilidade (Recall)} = \frac{\text{Verdadeiros Positivos (TP)}}{\text{Verdadeiros Positivos (TP)} + \text{Falsos Negativos (FN)}}$$

Em outras palavras, a sensibilidade responde à pergunta: "Quão bom o modelo é em detectar os verdadeiros positivos em relação a todos os exemplos positivos reais?"

Matriz de confusão – Acurácia

Acurácia : é uma métrica fundamental para avaliar a performance geral de um modelo de classificação. Ela mede a proporção de exemplos classificados corretamente em relação ao total de exemplos.

Matriz de confusão – Acurácia

A fórmula para calcular a acurácia é:

$$\text{Acurácia} = \frac{\text{Verdadeiros Positivos (TP)} + \text{Verdadeiros Negativos (TN)}}{\text{Total de Amostras}}$$

Em outras palavras, a acurácia responde à pergunta: "Qual a proporção de exemplos que o modelo classificou corretamente, independentemente da classe?"

- Verdadeiros Positivos (TP) são os casos em que o modelo previu corretamente a classe positiva.
- Verdadeiros Negativos (TN) são os casos em que o modelo previu corretamente a classe negativa.
- Total de Amostras é a soma de Verdadeiros Positivos, Falsos Positivos, Verdadeiros Negativos e Falsos Negativos.

Matriz de confusão – F1-Score

- F1 Score : também conhecido como pontuação F1, é uma métrica que combina precisão e sensibilidade (recall) em um único valor, proporcionando uma medida geral do desempenho de um modelo de classificação.

Matriz de confusão – F1-Score

O F1 Score é calculado pela média harmônica da precisão e da sensibilidade (recall). A média harmônica dá mais peso aos valores menores. A fórmula para calcular o F1 Score é:

$$\text{F1 Score} = 2 \times \frac{\text{Precisão} \times \text{Sensibilidade (Recall)}}{\text{Precisão} + \text{Sensibilidade (Recall)}}$$

Em outras palavras, o F1 Score é uma medida do equilíbrio entre precisão e sensibilidade. Ele é útil quando há um desequilíbrio entre as classes de interesse.

Um valor de F1 Score próximo de 1 indica um modelo com boa precisão e sensibilidade. Um valor de 0 indica um desempenho muito ruim.

O F1 Score é particularmente útil em problemas de classificação binária, onde há duas classes de interesse, mas também pode ser calculado para problemas de classificação multiclasse usando a média ponderada dos F1 Scores de cada classe.

Matriz de confusão – Geral

$$Accuracy = \frac{VN + VP}{VN + VP + FN + FP}$$

$$Precision = \frac{VP}{VP + FP}$$

$$Recall = \frac{VP}{VP + FN}$$

$$F1\ Score = 2.\frac{Precision\ .\ Recall}{Precision + Recall}$$

Underfitting

- Ocorre quando um modelo é muito simples para capturar os padrões dos dados de treino, resultando em um desempenho ruim tanto nos dados de treino quanto nos dados de teste..

Underfitting

- **Desempenho Ruim nos Dados de Treino**: O modelo não consegue capturar os padrões subjacentes dos dados de treino.
- **Desempenho Ruim nos Dados de Teste**: O modelo também falha ao generalizar para novos dados.
- **Simplicidade Excessiva**: O modelo é muito simples, com poucos parâmetros ou camadas, o que impede que ele aprenda adequadamente os padrões dos dados

Generalização vs Super Ajuste (*Overfitting*)

- Objetivo de todo classificador é criar modelos genéricos, que possam ser reutilizados em vários contexto.
- Um modelo super ajustado funciona bem com dados de treino, mas pode ter um desempenho péssimo quando submetido a dados de teste ou produção.
- Isso significa que o modelo aprendeu os detalhes e o ruído dos dados de treino, mas não consegue aplicar esse conhecimento a dados que nunca viu antes.

Generalização vs Super Ajuste (*Overfitting*)

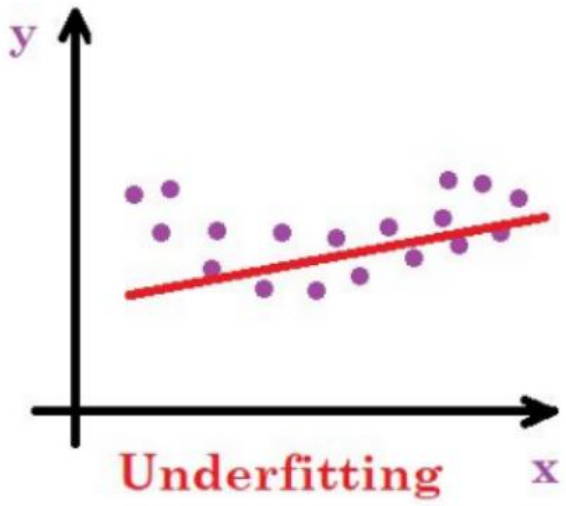
- **Desempenho Excelente nos Dados de Treino**: O modelo tem um erro muito baixo nos dados de treino.
- **Desempenho Ruim nos Dados de Teste**: O modelo apresenta um erro alto quando aplicado a novos dados.
- **Complexidade Excessiva**: O modelo é muito complexo, com muitos parâmetros ou camadas, o que permite que ele memorize os dados de treino

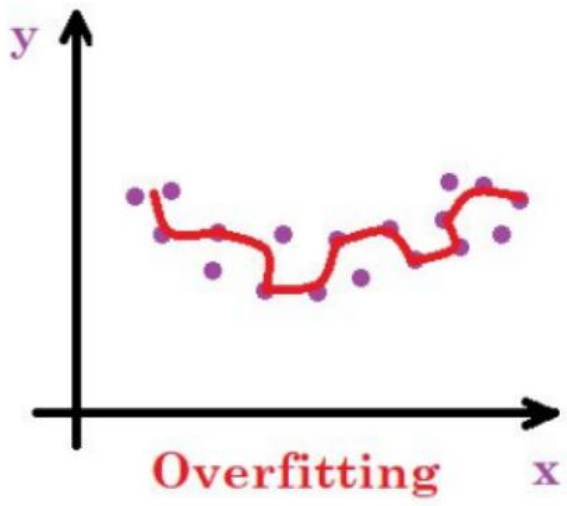
Underfitting X Overfitting

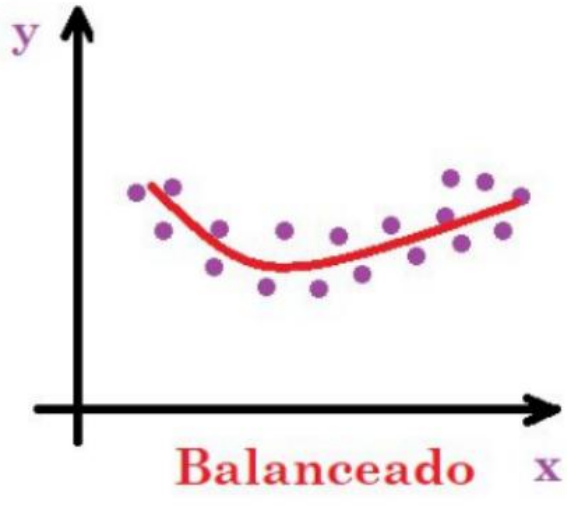
- **Overfitting**: Modelo muito complexo, bom desempenho nos dados de treino, mas ruim nos dados de teste.
- **Underfitting**: Modelo muito simples, desempenho ruim tanto nos dados de treino quanto nos dados de teste.

Underfitting X Overfitting

- **Underfitting**: A linha é quase reta e não segue a tendência dos pontos.
- **Overfitting** : A linha passa por todos os pontos, incluindo os outliers, resultando em uma curva muito complexa
- **Balanceado** : A linha segue a tendência geral dos pontos, ignorando o ruído e os outliers







Referências

- FREITAS, Rodrigo C. **A brief tutorial on using Python to make predictions - Breast Cancer Wisconsin (Diagnostic) Data Set**. 2017. Disponível em: <https://www.kaggle.com/rcfreitas/python-ml-breast-cancer-diagnostic-data-set/notebook>. Acesso em: 04 outubro 2018.
- HONDA, Hugo; FACURE, Matheus; YAOHAO, Peng. **Os Três Tipos de Aprendizado de Máquina**. 2017. Disponível em: <https://lamfo-unb.github.io/2017/07/27/tres-tipos-am/>. Acesso em: 02 agosto 2020.
- MORALES, Michele. **Como Construir um Classificador de Machine Learning em Python com Scikit-learn**. 2018. Disponível em: <https://www.digitalocean.com/community/tutorials/como-construir-um-classificador-de-machine-learning-em-python-com-scikit-learn-pt>. Acesso em: 04 outubro 2018.
- SANTANA, Rodrigo. **Tipos de Aprendizado de Máquina e a Historia dos Criadores de Vacas**. 2018. Disponível em: <http://minerandodados.com.br/index.php/2018/03/20/tipos-de-aprendizado-de-maquina/>. Acesso em: 04 outubro 2018.

OBRIGADO

Wanderlan Carvalho

Professora de Ensino Superior

E-mail:
032000093@prof.uninorte.com.br