



**UNIVERSIDADE FEDERAL DE CAMPINA GRANDE**  
INICIAÇÃO CIENTÍFICA E TECNOLÓGICA  
UNIDADE ACADÊMICA DE ENGENHARIA ELÉTRICA

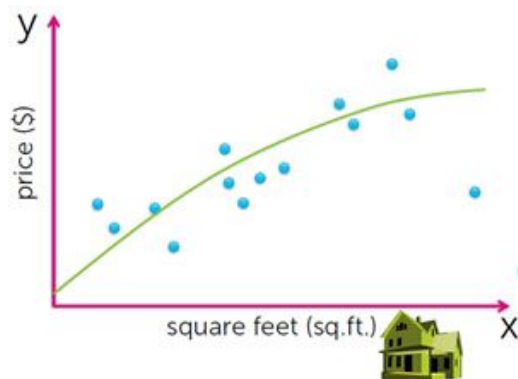
# Avaliação de modelos em aprendizado de máquina

**Reunião 20/11/2020**

# Modelo com muitos erros na predição

Vamos supor que foi implementado um modelo de regressão linear com regularização para prever o preço de casas em uma determinada cidade:

$$J(\theta) = \frac{1}{2m} \left[ \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^m \theta_j^2 \right]$$



# Modelo com muitos erros na predição

Ao avaliar a predição do algoritmo nos dados de treinamento, foi observado que o modelo estipulado estaria passível a diversos erros de predição. Algumas medidas foram pensadas:

- Adicionar mais dados de treinamento;
- Tentar diminuir o número de características dos dados;
- Adicionar um hipótese de polinômios com grau mais elevados;
- Aumentar o valor de  $\lambda$  na regularização;
- Diminuir o valor de  $\lambda$  na regularização;

Todas essas tentativas podem ser eficazes, embora gaste muito tempo para analisá-las.

# Diagnóstico do Aprendizado de Máquina

- Um teste que possibilitar a obtenção de uma visão do que está ou não funcionando com um algoritmo de aprendizagem, visando obter orientação sobre a melhor forma de melhorar o desempenho do algoritmo.
- A implementação de diagnósticos no modelo a ser estipulado pode levar algum tempo, mas o resultado tende a ser valioso para boas previsões.

Tais ideias partem da premissa de que não se pode lançar um modelo sem antes garantir a sua capacidade de generalização.

# Abordagem do AM



Prepare Data



Engineer Features



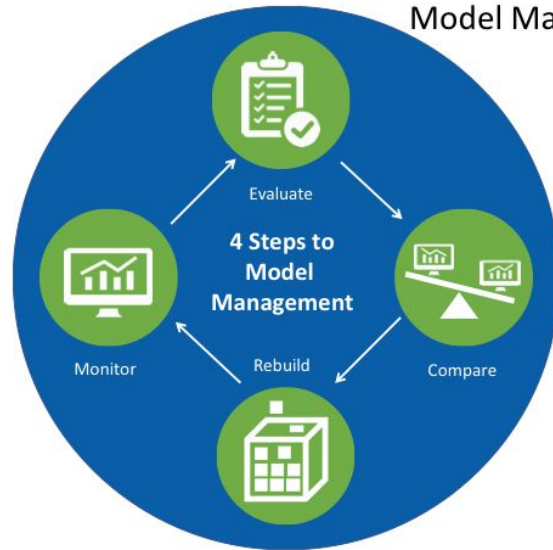
Train, Build &  
Test Models

Aplicar técnicas de AM para se aprofundar em grandes quantidades de dados pode ajudar na descoberta de padrões que não eram aparentes. Isto é chamado de mineração de dados.

Deploy Best  
Performing Model

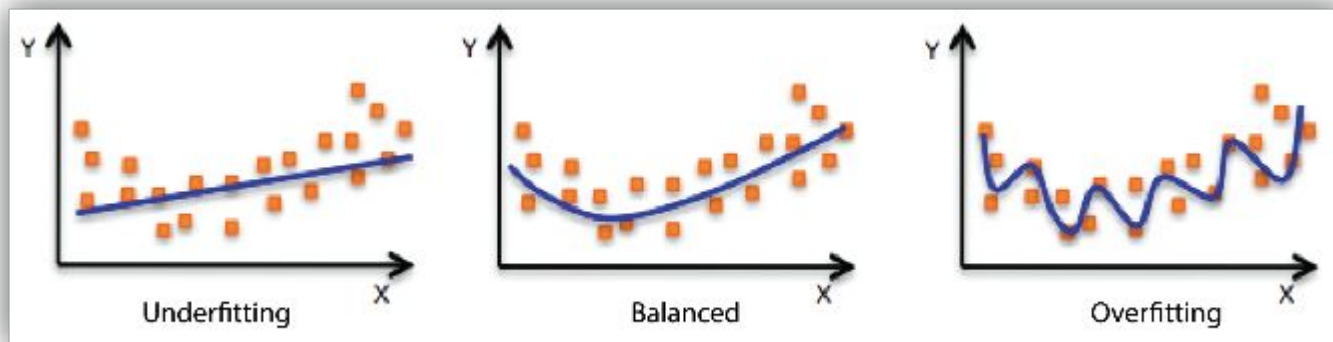


Machine Learning  
Model Management



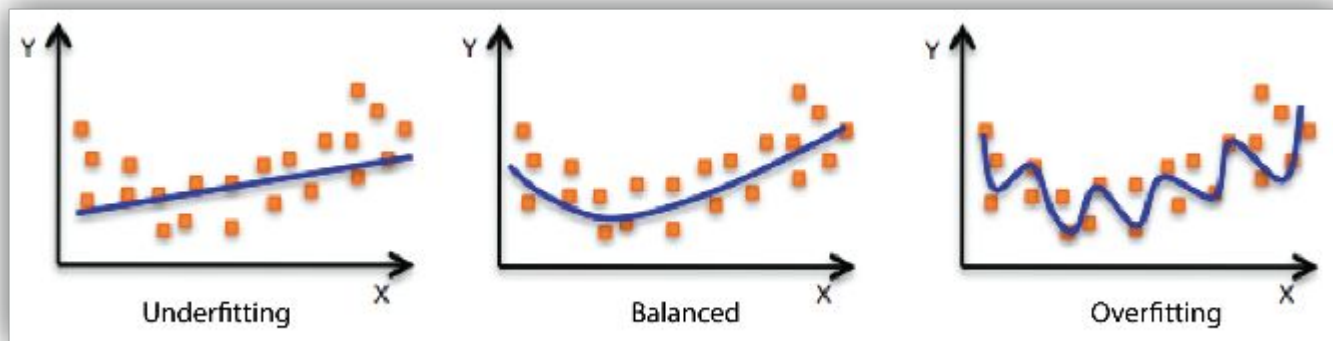
# Avaliando a Hipótese

- **Underfitting** = o modelo não se ajustou muito bem aos dados. Consequentemente, é passível de avaliar incorretamente dados de treinamento e outros novos dados.



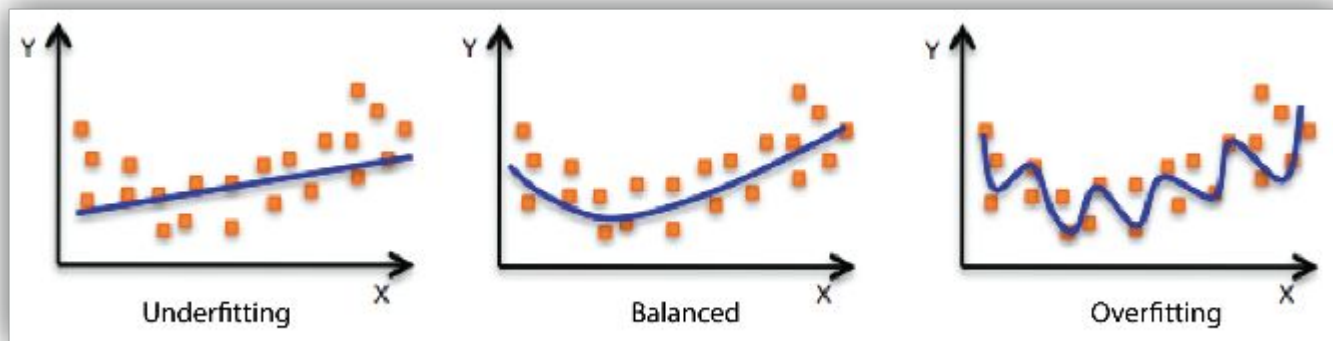
# Avaliando a Hipótese

- **Ajuste perfeito** = O modelo se ajustou bem aos dados de modo que ele consiga avaliar concisamente os dados de treinamento e outros novos dados. Ou seja, o modelo tem capacidade de generalização.



# Avaliando a Hipótese

- **Overfitting** = O modelo apenas decorou os dados. Houve um sobreajuste enorme aos dados de treinamento, tornando-o sensível a novos dados que destoam do que foi decorado.





# Avaliando a Hipótese

## Undefitting

- $J(\theta)$  alto e  $J_{test}(\theta)$  alto;

## Ajuste Perfeito

- $J(\theta)$  consistente e  $J_{test}(\theta)$  consistente;

## Overfitting

- $J(\theta)$  baixo e  $J_{test}(\theta)$  alto;

É importante ter cuidado com avaliações parciais sobre um modelo de um algoritmo de aprendizado de máquina.



# Avaliando a Hipótese

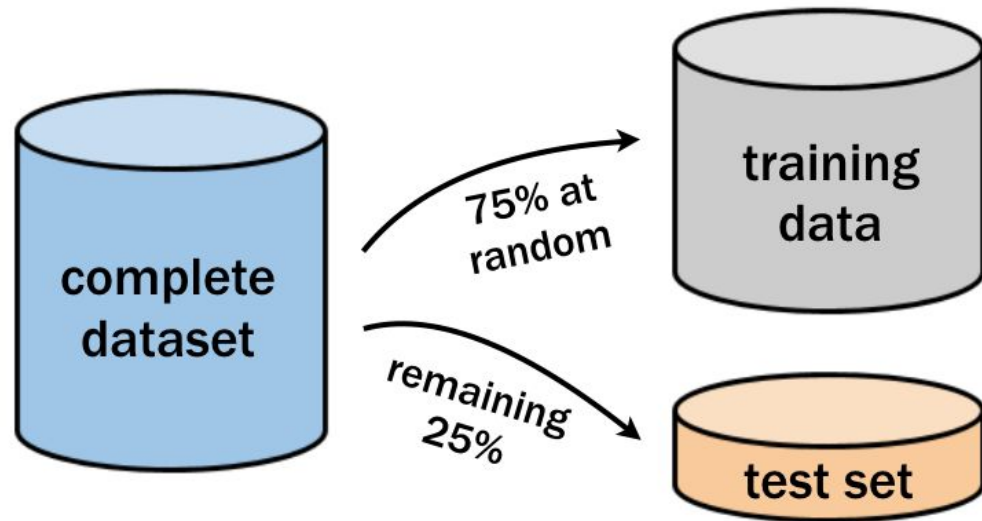
Como seria possível avaliar o ajuste dos parâmetros do modelo em relação a sua capacidade de generalização?

Divisão do dataset completo em:

- **Dados de treinamento;**
- **Dados de teste;**

Aprendizagem Supervisionada:

**(Dados Previsores, Rótulos)**



# Procedimentos em regressão linear/polinomial

- Realizar a minimização na função de custo  $J(\theta)$  por meio do ajuste de parâmetros utilizando os dados de treinamento;
- Computar o erro quadrático médio vinculado aos dados de teste;

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} (h_{\theta}(x_{test}^{(i)}) - y_{test}^{(i)})^2$$

- Avaliar o ajuste dos dados em relação a capacidade de generalização;

# Procedimentos em regressão logística

- Realizar a minimização na função de custo  $J(\theta)$  por meio do ajuste de parâmetros utilizando os dados de treinamento;
- Computar o erro quadrático médio vinculado aos dados de teste;

$$J_{test}(\theta) = -\frac{1}{m_{test}} \sum_{i=1}^{m_{test}} y_{test}^{(i)} \log h_{\theta}(x_{test}^{(i)}) + (1 - y_{test}^{(i)}) \log h_{\theta}(x_{test}^{(i)})$$

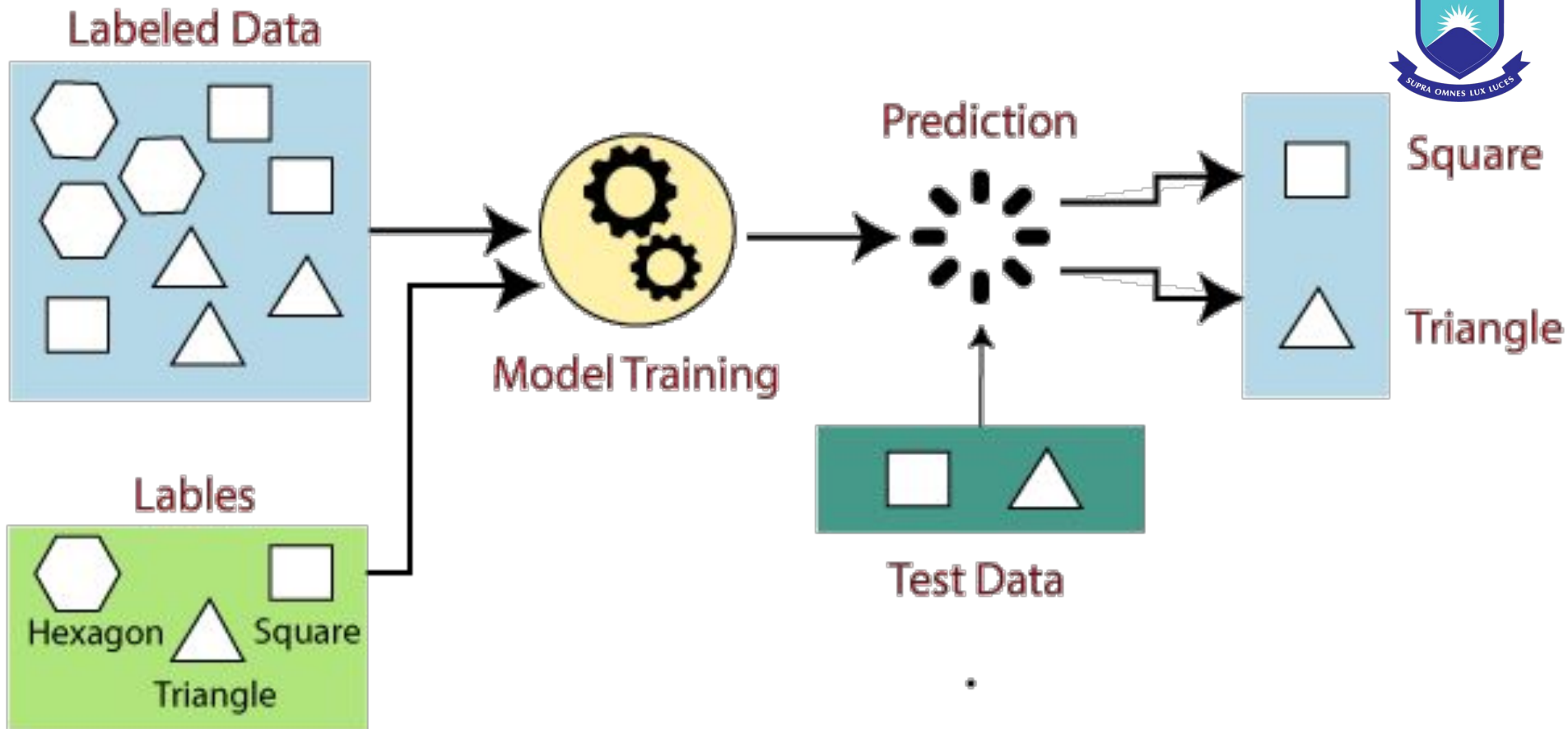
- Avaliar o ajuste dos dados em relação a capacidade de generalização;

# Erro de Classificação errada

$$err(h_{\Theta}(x), y) = \begin{cases} 1 & \text{if } h_{\Theta}(x) \geq 0.5 \text{ and } y = 0 \text{ or } h_{\Theta}(x) < 0.5 \text{ and } y = 1 \\ 0 & \text{otherwise} \end{cases}$$

- Em casos de classificação binária, podemos computar o chamado erro de classificação errada, usando como artifício a análise das rotulagens corretas e incorretas da predição.

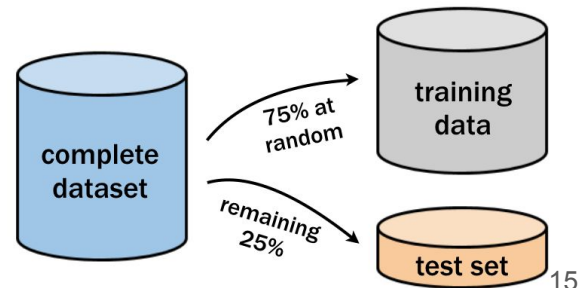
$$\text{Test Error} = \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} err(h_{\Theta}(x_{test}^{(i)}), y_{test}^{(i)})$$



# Dados de Treinamento, Validação e Teste

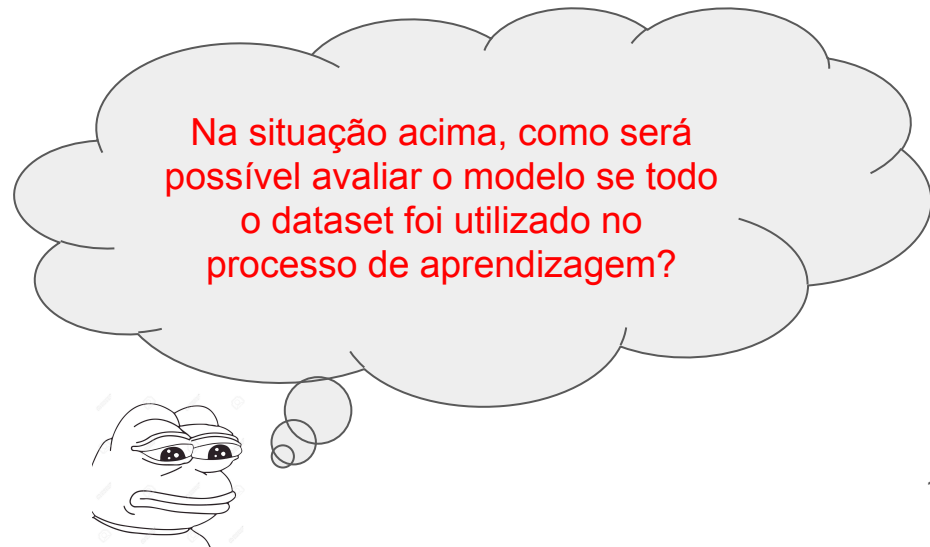
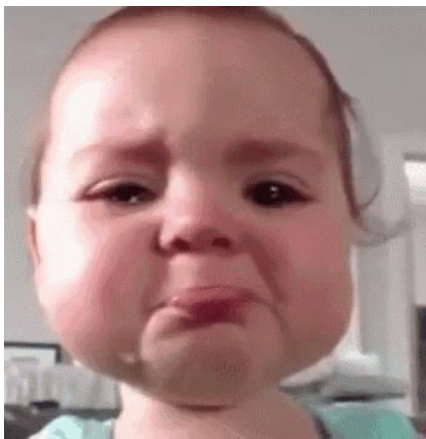
Digamos que em um problema de regressão polinomial, deseja-se encontrar qual ordem  $n$  de um polinômio consegue ter predições mais generalistas em relação a novos dados.

Uma alternativa é minimizar a função de custo através dos ajustes dos parâmetros e computar o erro de predição com os dados de treinamento. Desse modo, podemos encontrar um  $n$  grau de polinômio vinculado a hipótese que satisfaça o meu problema.



# Dados de Treinamento, Validação e Teste

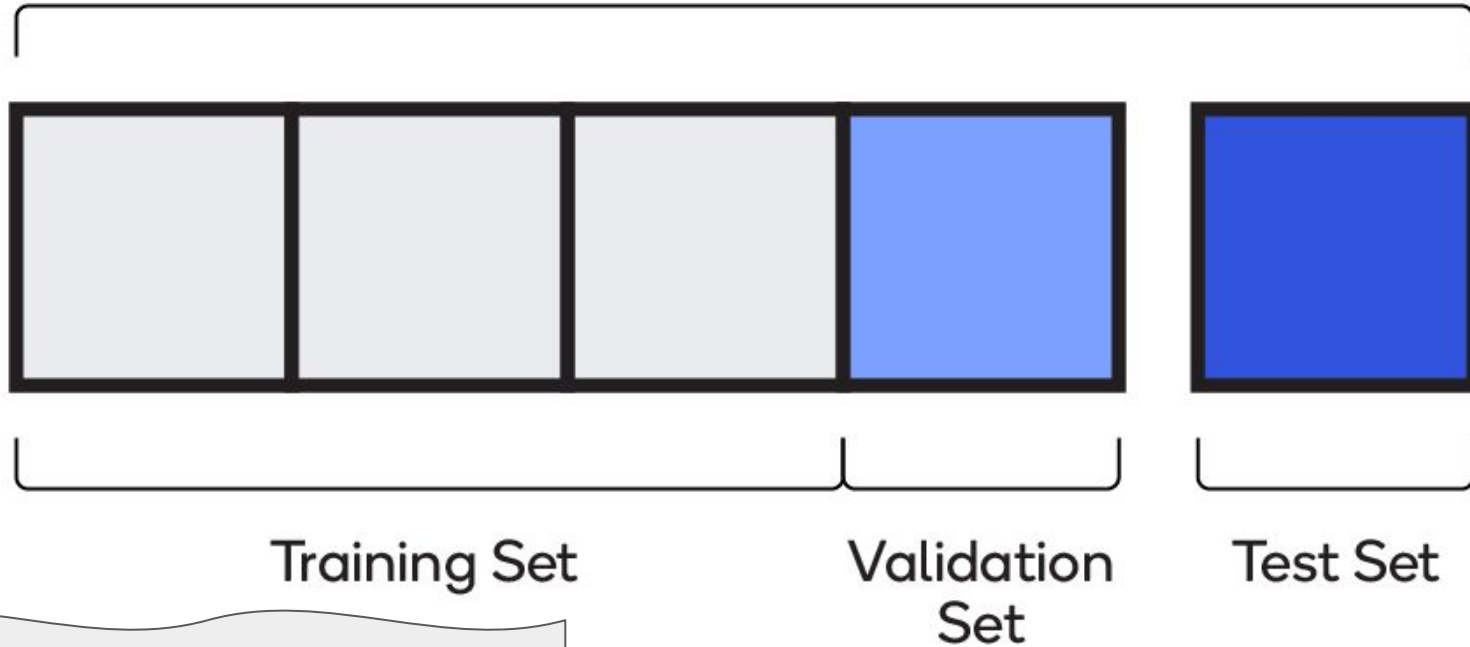
Pensando mais sobre o problema anterior, é possível perceber que essa não foi uma escolha precisa, pois os dados de teste, que originalmente deveriam se manter intactos do processo de aprendizagem do algoritmo, foram constantemente utilizados na busca de uma melhor hipótese.





# Dados de Treinamento, Validação e Teste

Dataset  $D$



Mantém os dados de teste em segurança para cumprir o seu propósito.

# Dados de Treinamento, Validação e Teste

## Dados de Treinamento

- Utilizado para realizar o processo de minimização da função de custo através do ajuste dos parâmetros;

## Dados de Validação

- Utilizado para computar o erro na predição das diversas hipóteses com polinômios de grau  $n$  (escolha do ajuste ideal);

## Dados de Teste

- Verifica a capacidade de generalização do modelo usando a hipótese obtida;

# Matriz de Confusão

		Valor Previsto	
		Positivo	Negativo
Valor Verdadeiro	Positivo	<b>TP</b> <b>Verdadeiro Positivo</b>	<b>FN</b> <b>Falso Negativo</b>
	Negativo	<b>FP</b> <b>Falso Positivo</b>	<b>TN</b> <b>Verdadeiro Negativo</b>

# Matriz de Confusão

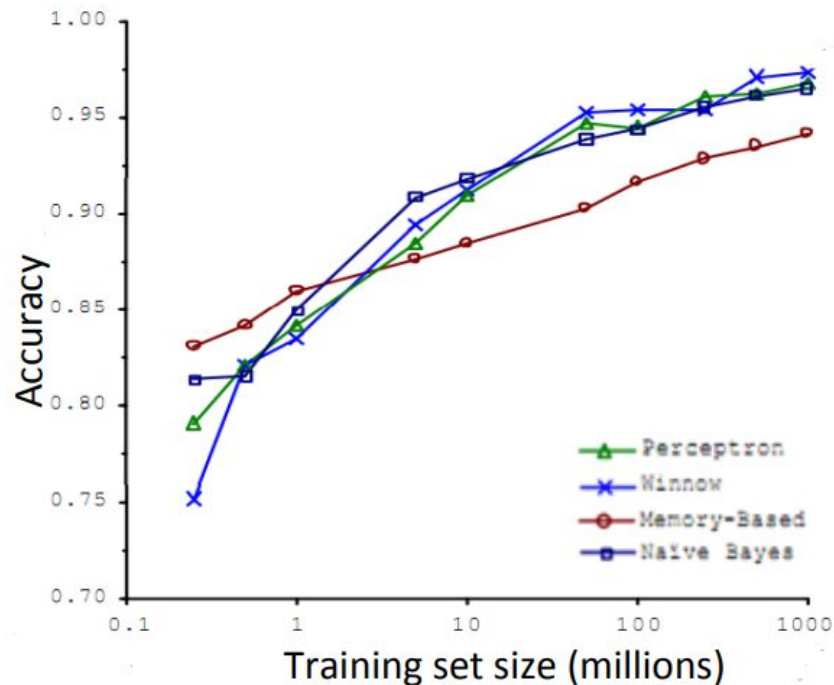
		Classe esperada	
		Gato	Não é gato
Classe prevista	Gato	25	10
	Não é gato	25	40

# Matriz de Confusão

		Classe esperada	
		Gato	Não é gato
Classe prevista	Gato	<b>25</b> Verdadeiro Positivo	<b>10</b> Falso Positivo
	Não é gato	<b>25</b> Falso Negativo	<b>40</b> Verdadeiro Negativo

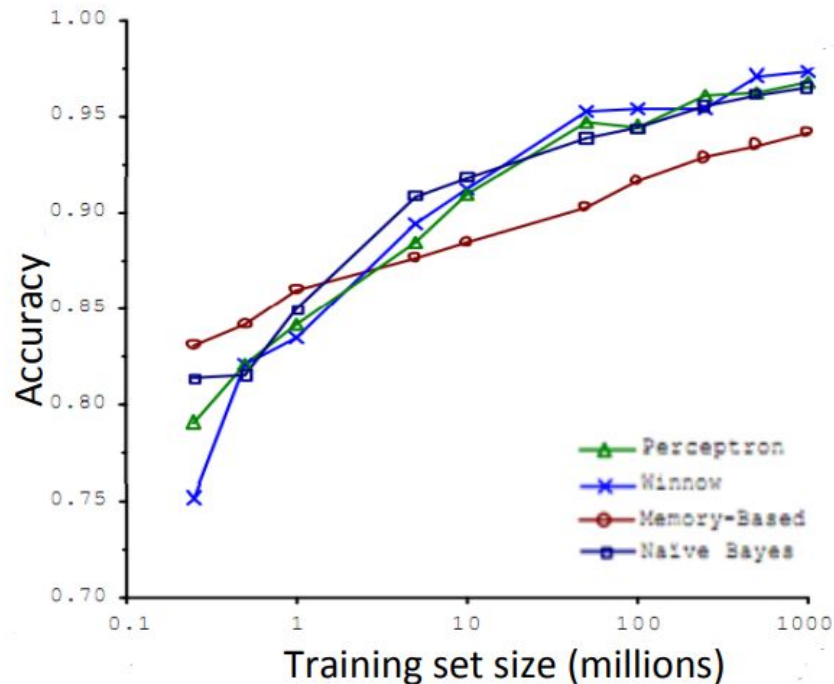
# Quantidade de Dados de Treinamento

- Mais chances de aprendizado;
- É uma boa base de dados para treinamento, pois possui dados mais generalistas e consistentes;
- É comum ter a necessidade de eliminar valores inconsistentes, faltantes e que fogem radicalmente dos padrões (outliers);

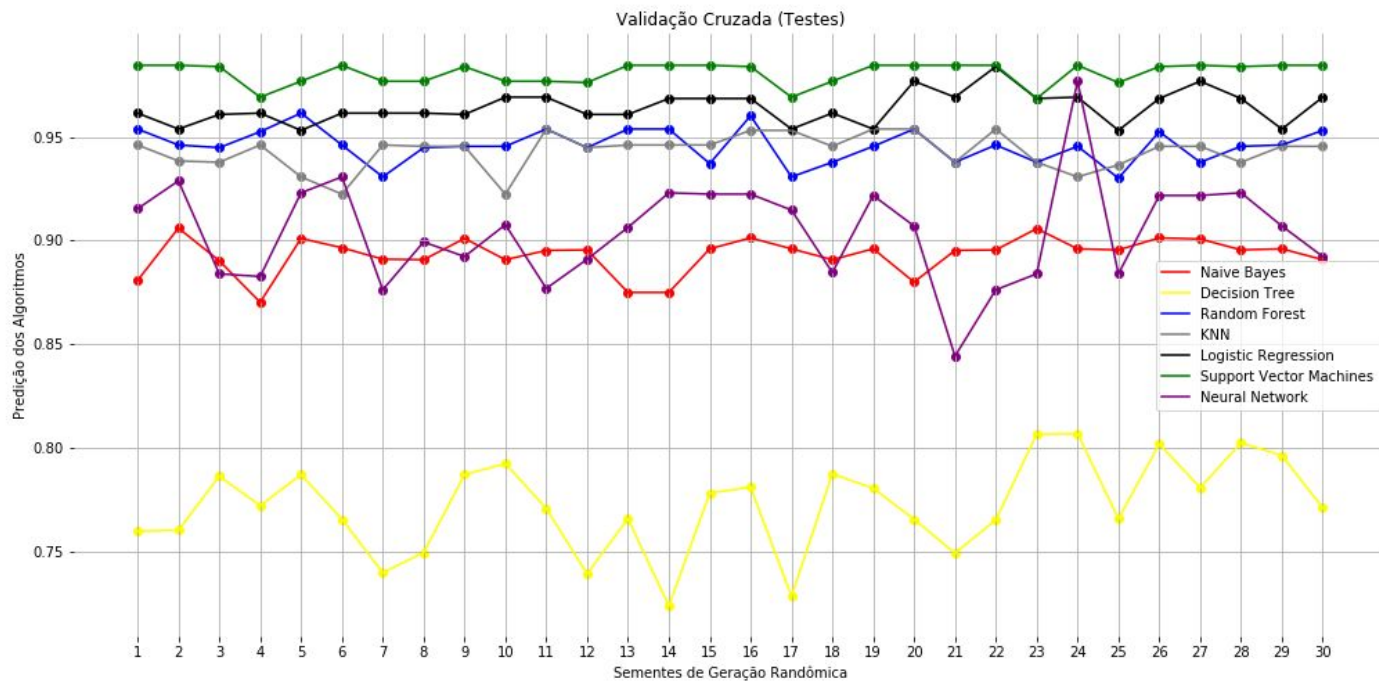


# Testando novos algoritmos

- É comum nos processos de aprendizagem de máquina testar diferentes algoritmos em um mesmo problema e analisar qual terá o melhor desempenho;



# Testando novos algoritmos





Continua...

