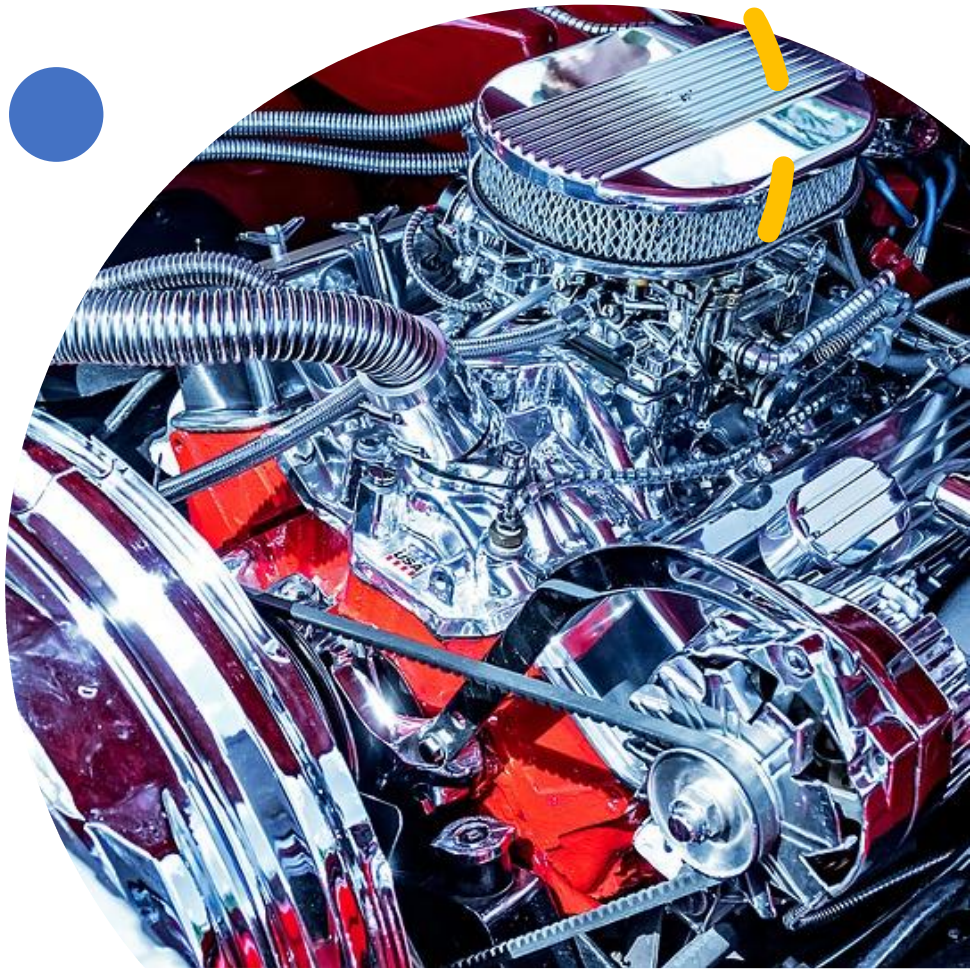
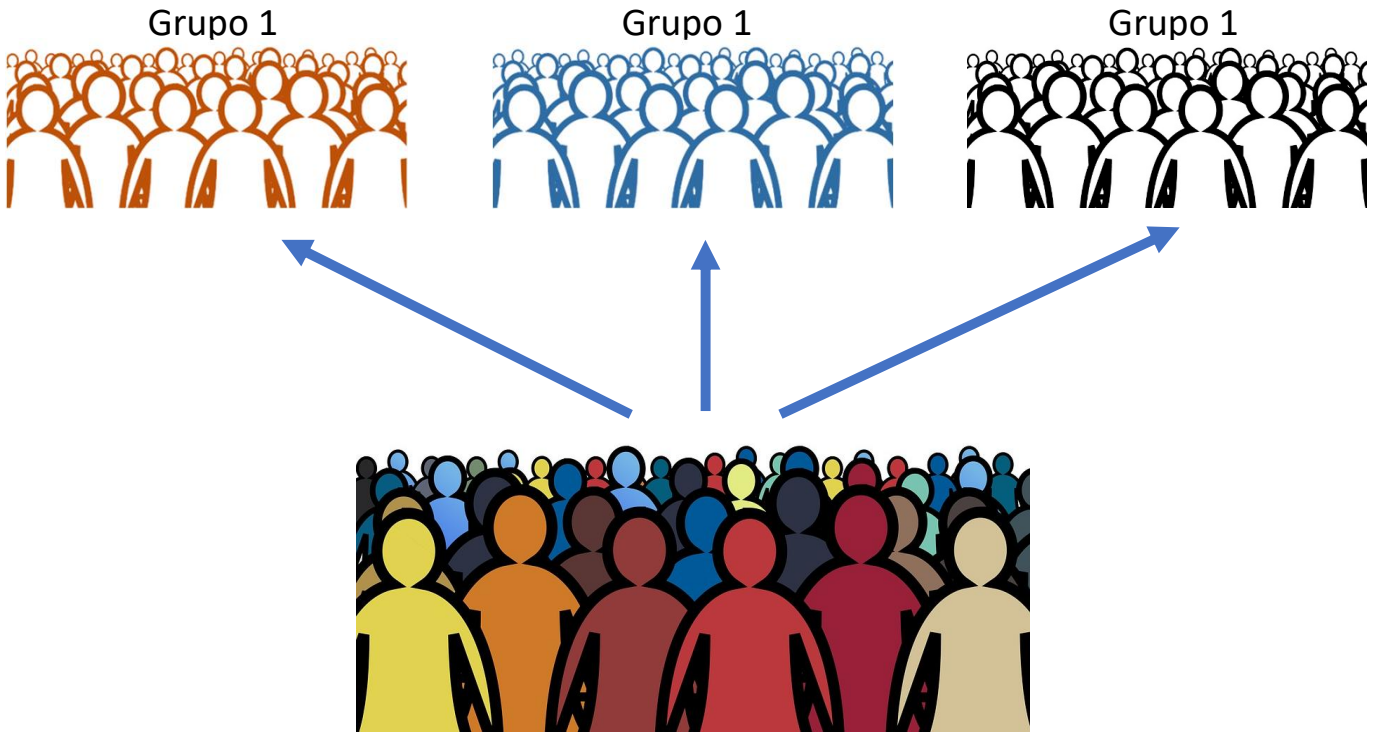
Regressão

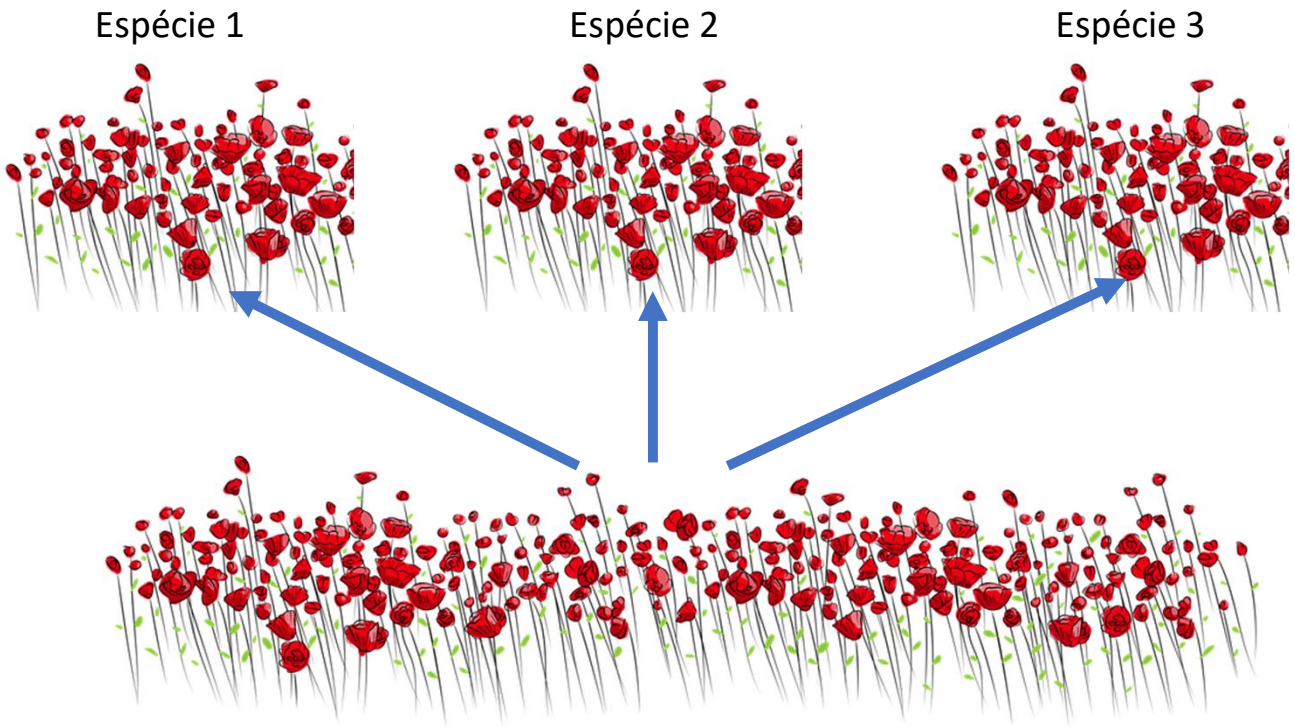
- Qual a Potência?



Agrupamentos



Agrupamentos



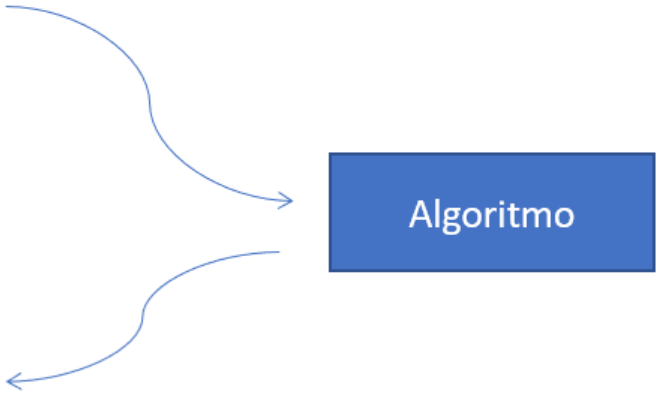
Sistemas de Recomendação

Qual a confiança e precisão?

Modelo

Dados Históricos de Concessão de Crédito

Idade	Pagou
18	Não
46	Sim
34	Sim
21	Não
37	Não
...	



Modelo Construído

Idade	Bom Pagador
18~22	Não
23~35	Sim
36~45	Não
45~65	Sim

Novo cliente com 37 anos. Bom ou mal pagador?

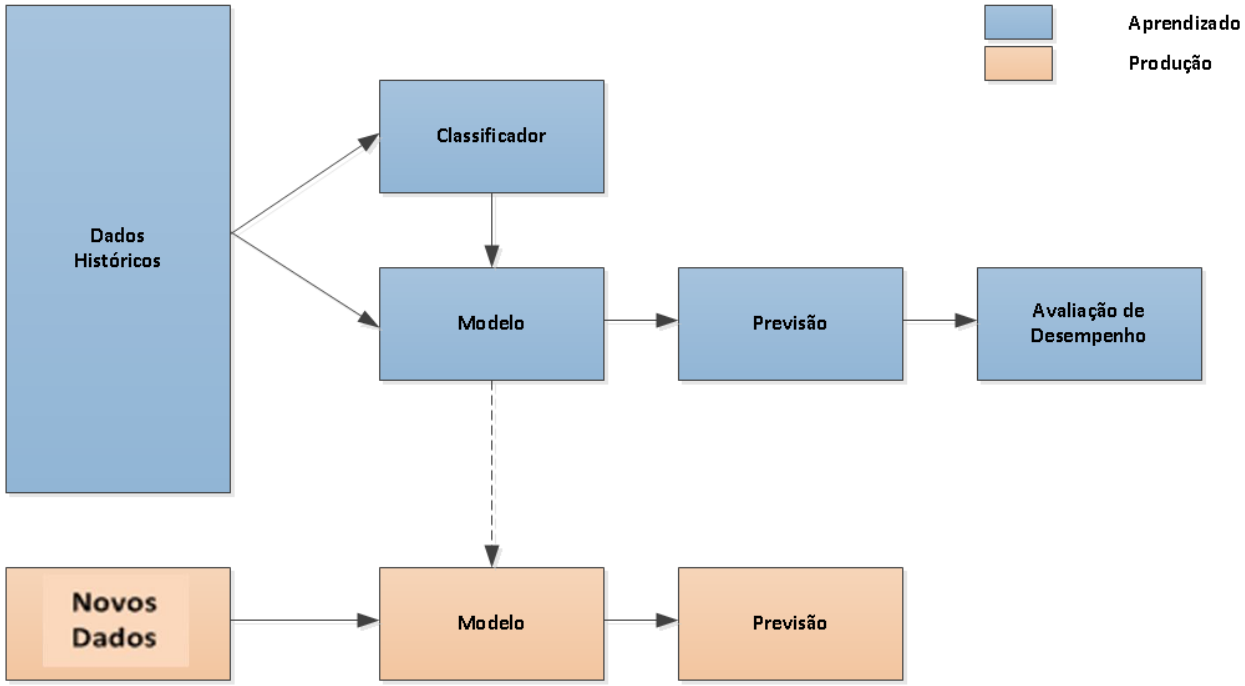
Medir o Desempenho do Modelo

Treino: Algoritmo processa dados e cria modelo

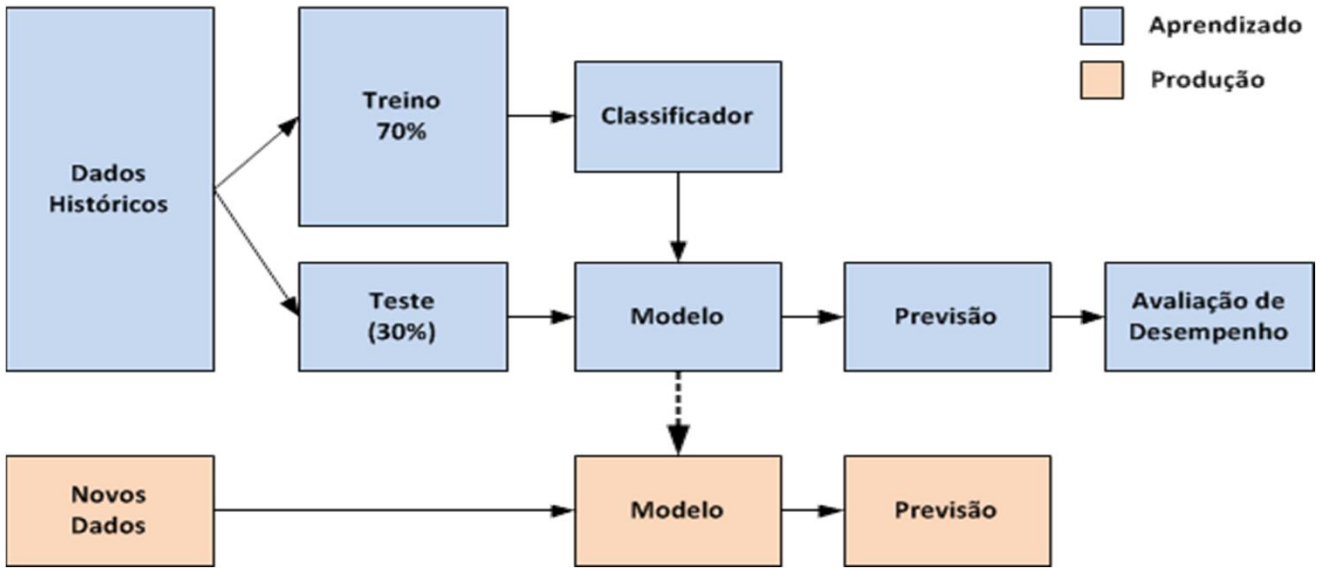
Teste: Dados são submetidos ao modelo e se mede a precisão

1. Mesmo conjunto de dados
2. Hold out
3. Sub-amostragem Aleatória
4. Validação Cruzada

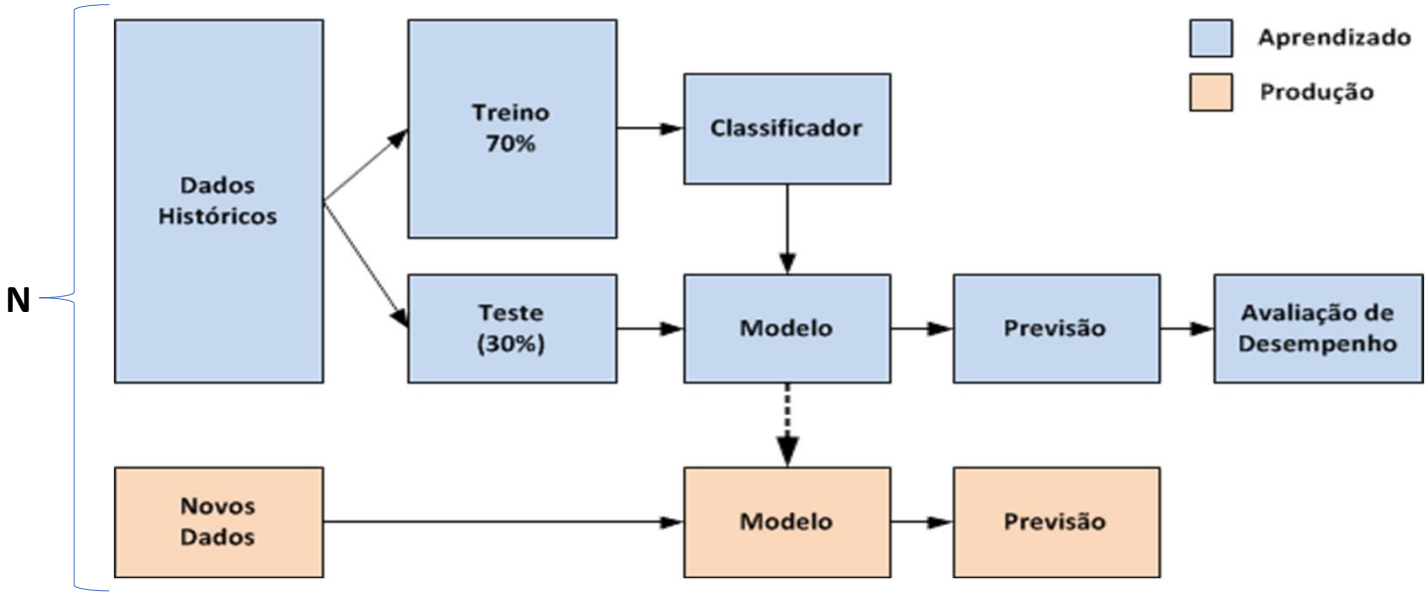
1 - Usando mesmo conjunto de dados



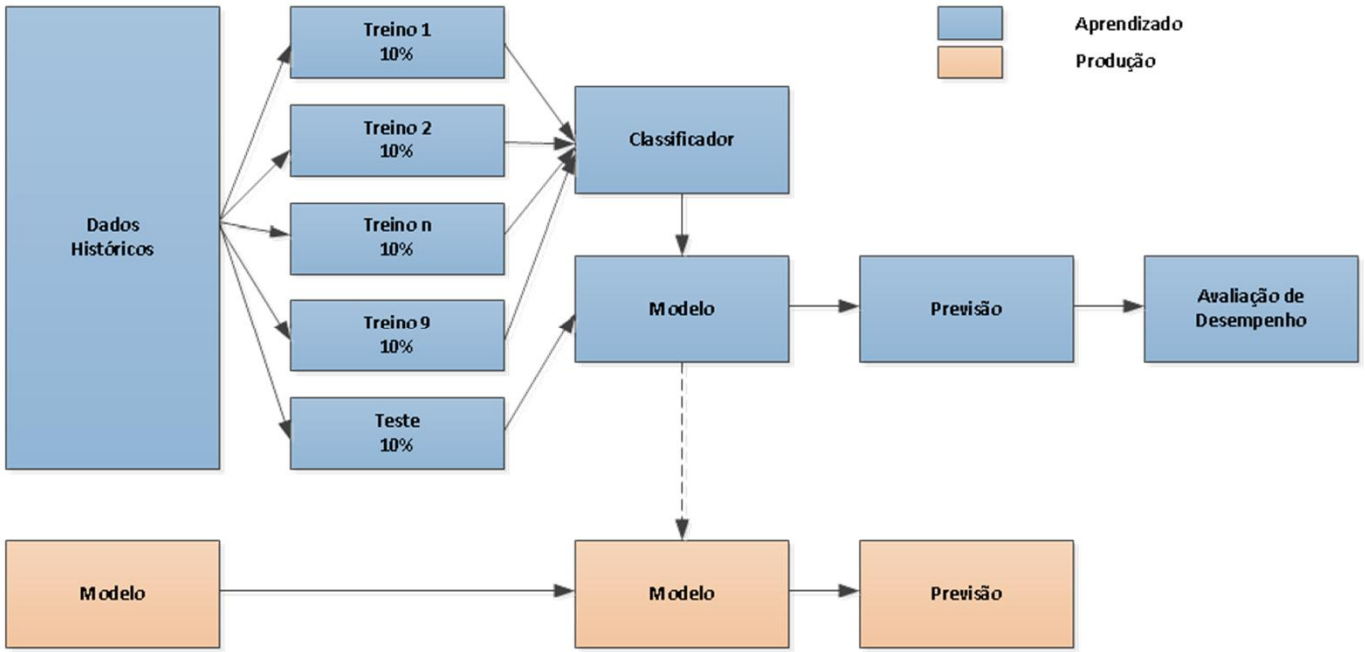
2 – Hold out


Aprendizado
Produção
Dados Históricos
Treino 70%
Classificador
Teste (30%)
Modelo
Previsão
Avaliação de Desempenho
Novos Dados
Modelo
Previsão

3 – Sub-amostragem Aleatória



4 – Validação Cruzada



Matriz de Confusão

Idade	Pagou	Classificação
18	Não	Não
46	Sim	Sim
34	Sim	Não
21	Não	Sim
37	Não	Não
...		

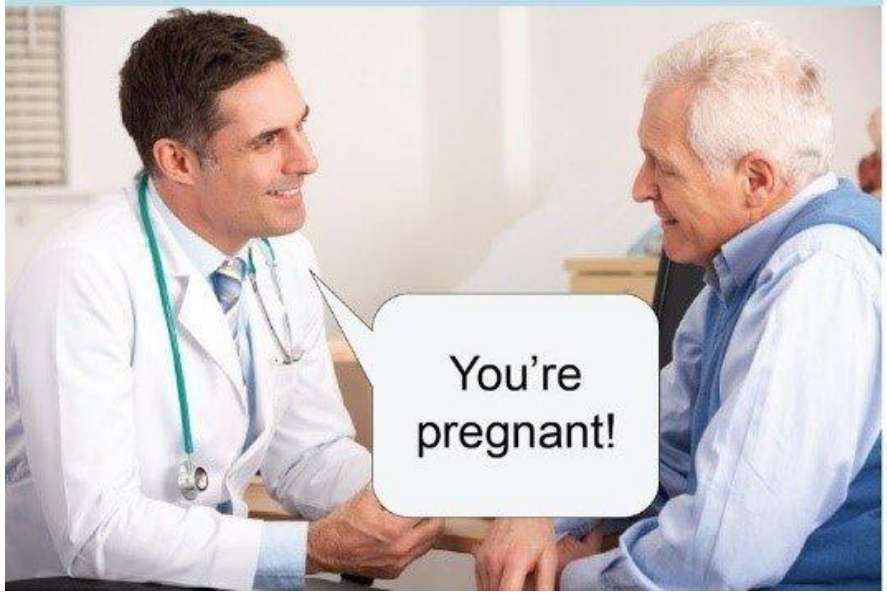
		Classificação	
		Sim	Não
Dados	Sim	1	1
	Não	1	2

Matriz de Confusão

		Classificação	
		Sim	Não
Dados	Sim	1	1
	Não	1	2

Verdadeiros Positivos	Falsos Negativos
Falsos Positivos	Verdadeiros Negativos

Type I Error



Type II Error

Generalização Versus Super Ajuste Versus Sub Ajuste

- O objetivo de todo classificador é criar modelos genéricos
- O modelo super ajustado funciona bem com dados de treino, mas tem o desempenho pobre em dados de teste ou de produção.

Genérico

Super/
Sub Ajustado

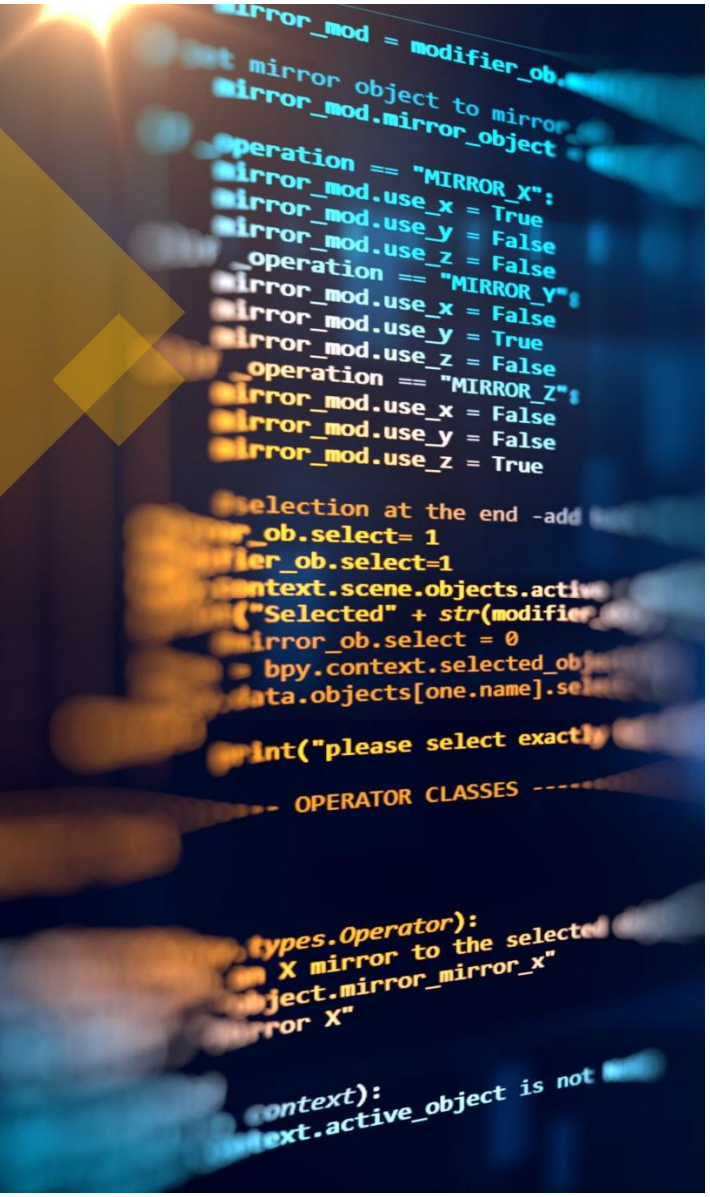
Generalização Versus Superajuste Versus Sub-ajuste

Explicados através do estudo para uma prova

Ana: Reprovada
Motivo: Superajuste
Estudou apenas através de provas anteriores. Seu aprendizado foi construído de forma a responder apenas as questões daquelas provas.

Carlos: Reprovado
Motivo: Sub-ajuste
Estudou apenas através de resumos e dicas. Seu método não foi capaz de aprender o conteúdo de forma genérica e ampla.

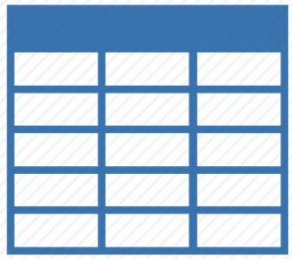
Maria: Aprovada
Motivo: Generalização
Estudou de forma ampla toda o conteúdo. Seu método foi capaz de criar um aprendizado <u>bom para qualquer prova.</u>



Causas de Super Ajuste

- Dados não representativos
- Dados não significativos (poucos)
- Forma de treinamento
- Classe rara
- Modelo incorreto

Problema da Classe Rara / Não balanceado



1 milhões de transações

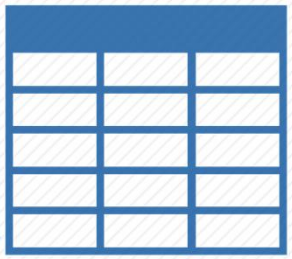
5% fraudulentas

Amostragem Aleatória sem reposição: 1%

10.000

9.900 - Legítimas

100 - Fraudulentas



1 milhões de transações

5% fraudulentas

Amostragem Aleatória sem reposição:

0,5% - Legítimas

0,5% - Fraudulentas

10.000

5.000 - Legítimas

5.000 - Fraudulentas

Problemas de Atributos Desconhecidos

- No treino: regiões “Sul”, “Sudeste”, “Centro-Oeste” e “Norte”
- Na produção: região “Nordeste”

Métricas para Classificação

		Treino	
		Aprovado	Reprovado
Originais	Aprovado	1004	208
	Reprovado	202	294

Métrica	Fórmula	Descrição	Cálculo
Acertos	(VP+VN)/Total	Total de Acertos	75,99
Erros	(FP+FN)/Total	Total de Erros	24,00
Precisão	VP/(VP+FP)	Quantos registros de fato são positivos	83,25
Lembrança ou Positivos Verdadeiros	VP/(VP+FN)	Positivos corretamente previstos	82,83
Negativos Verdadeiros	VN/(VN+FP)	Total de Negativos Falsos	59,27
Positivos Falsos	FP/(VN+FP)	Total de Positivos Falsos	40,72
Negativos Falsos	FN/(VP+FN)	Total de Negativos Falsos	17,16