FORMAÇÃO CIENTISTA DE DADOS

CLASSIFICAÇÃO

MATRIZ DE CONFUSÃO

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

		Sim	Não	Classificação
Dados	Sim	1	1	
	Não	1	2	

MATRIZ DE CONFUSÃO

		Sim	Não	Classificação
Dados	Sim	1	1	
	Não	1	2	

Verdadeiros Positivos	Falsos Negativos
Falsos Positivos	Verdadeiros Negativos

GENERALIZAÇÃO VERSUS SUPER 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 Ajustado

CAUSAS DE SUPER AJUSTE

- DADOS NÃO REPRESENTATIVOS
- DADOS NÃO SIGNIFICATIVOS (POUCOS)
- FORMA DE TREINAMENTO
- CLASSE RARA (DESCRITA EM SEGUIDA)
- MODELO INCORRETO

PROBLEMA DA CLASSE RARA

- TRANSAÇÕES DE FRAUDE: A FRAUDE É UMA CLASSE RARA
- O MODELO PODE TER DIFICULDADE DE APRENDER UMA CLASSE RARA
- SOLUÇÃO: ESTRATIFICAÇÃO

PROBLEMAS DE ATRIBUTOS DESCONHECIDOS

- NO TREINO: REGIÕES “SUL”, “SUDESTE”, “CENTRO-OESTE” E “NORTE”
- NA PRODUÇÃO: REGIÃO “NORDESTE”

- COMO CLASSIFICAR?