# 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 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.
- O modelo sub ajustado não consegue boas taxas de previsão. Ele não foi capaz de capturar as características do negócio para o modelo

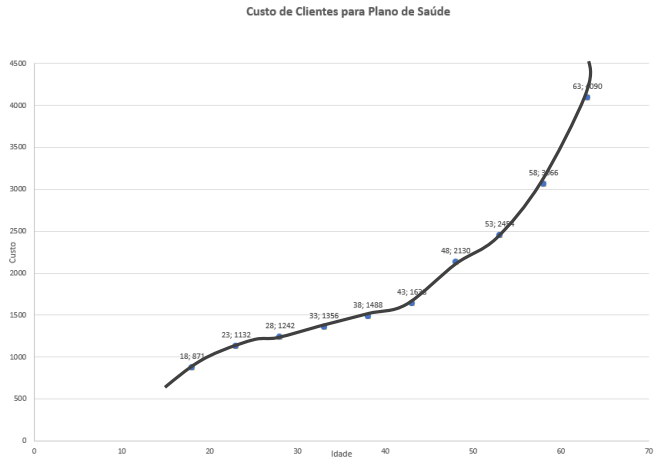
Genérico Super Ajustado Sub Ajustado

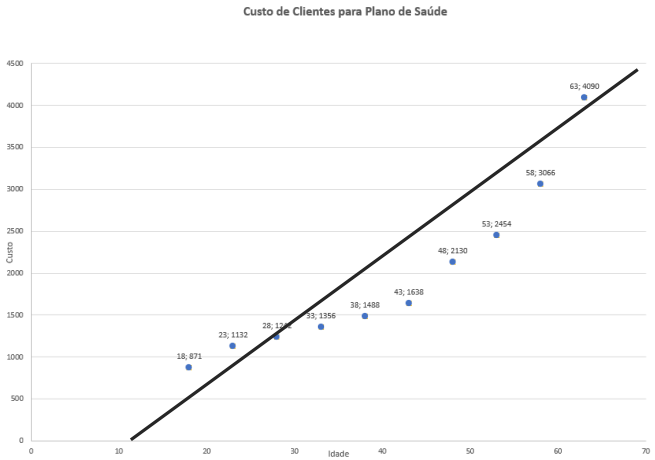
# Causas de Super/ Sub Ajuste

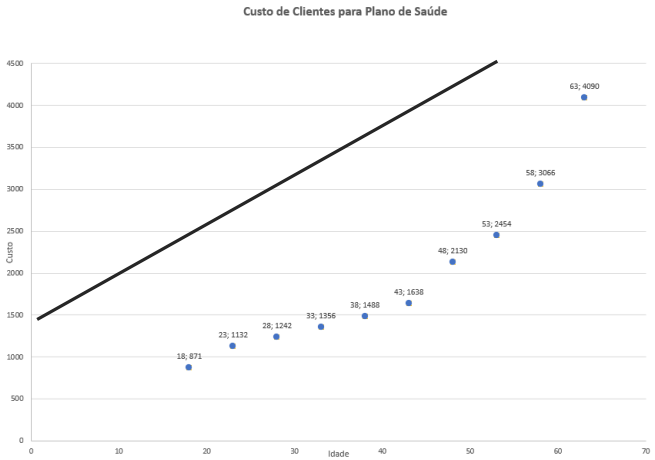
- Dados não representativos
- Dados não significativos (poucos)
- Forma de treinamento
- Classe rara (descrita em seguida)
- Modelo incorreto

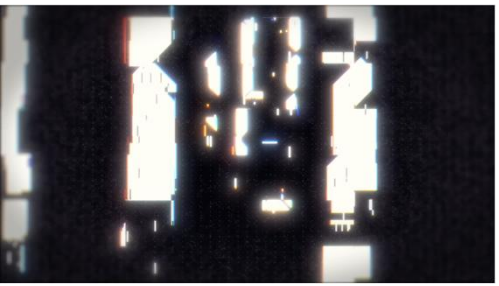
# Problema da Classe Rara

- Transações de Fraude: a fraude é uma classe rara
- O modelo pode ter dificuldade de aprender uma classe rara
- Solução: estratificação









# Qual é um bom modelo?

TESTE

☐ A

☑ B

☐ C

PROF.
FERNANDO
AMARAL
www.datascientist.com.br