Case 04: Modelos preditivos

Objetivos de aprendizagem de hoje

Contextualizar sobre a aplicação de jurimetria em empresas

Conhecer como funcionam e para que servem modelos preditivos

Conhecer o funcionamento geral de dashboards

Guia para os slides

Slides sobre o case: **marca azul**

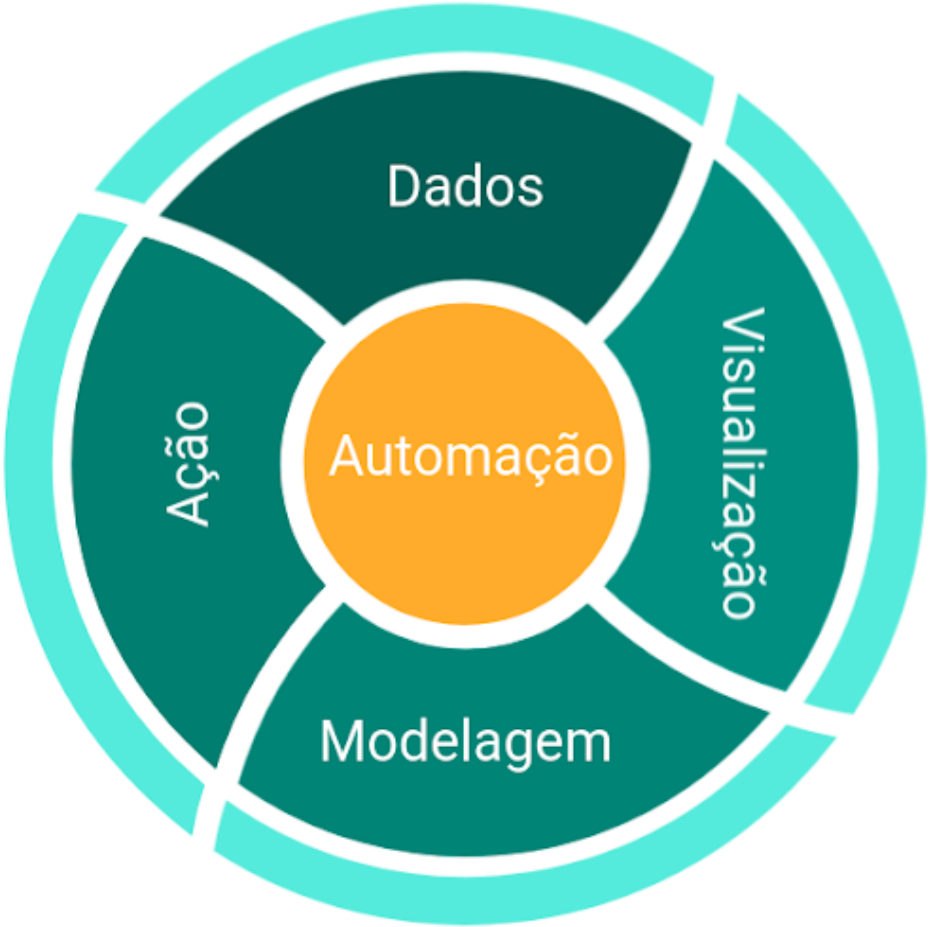
Slides sobre pesquisa/ciência: **marca verde**

Slides sobre estatística: **marca rosa**

Lawtechs

- AB2L surgiu em 2016
- Brasil apresenta um dos maiores mercados mercados

Ciclo de maturidade de dados



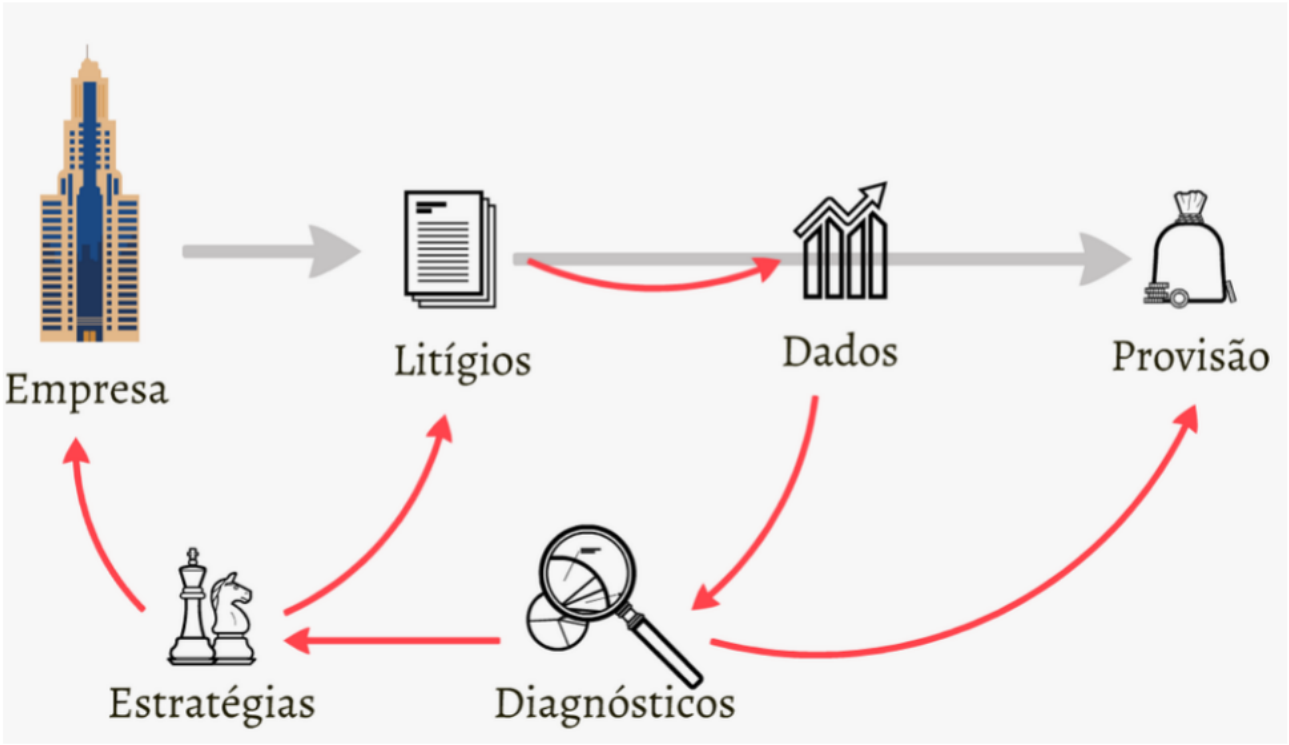
Cenário atual

- Por conta da dificuldade de acesso aos dados, ainda existe oportunidade para empresas especializadas em **extração de dados**.
- A maior parte das aplicações que vimos hoje são relacionadas a **visualização de dados**, com algumas aplicações pontuais de **modelagem**.
- A tomada de **decisão baseada em dados** ainda é bastante incipente no mercado atual (mesmo que as empresas não falem isso explicitamente).

> Fake it until you make it.
>
> Simon & Garfunkel (?)

Jurimetria em Empresas

invertendo a imagem do setor jurídico



Sobre o case

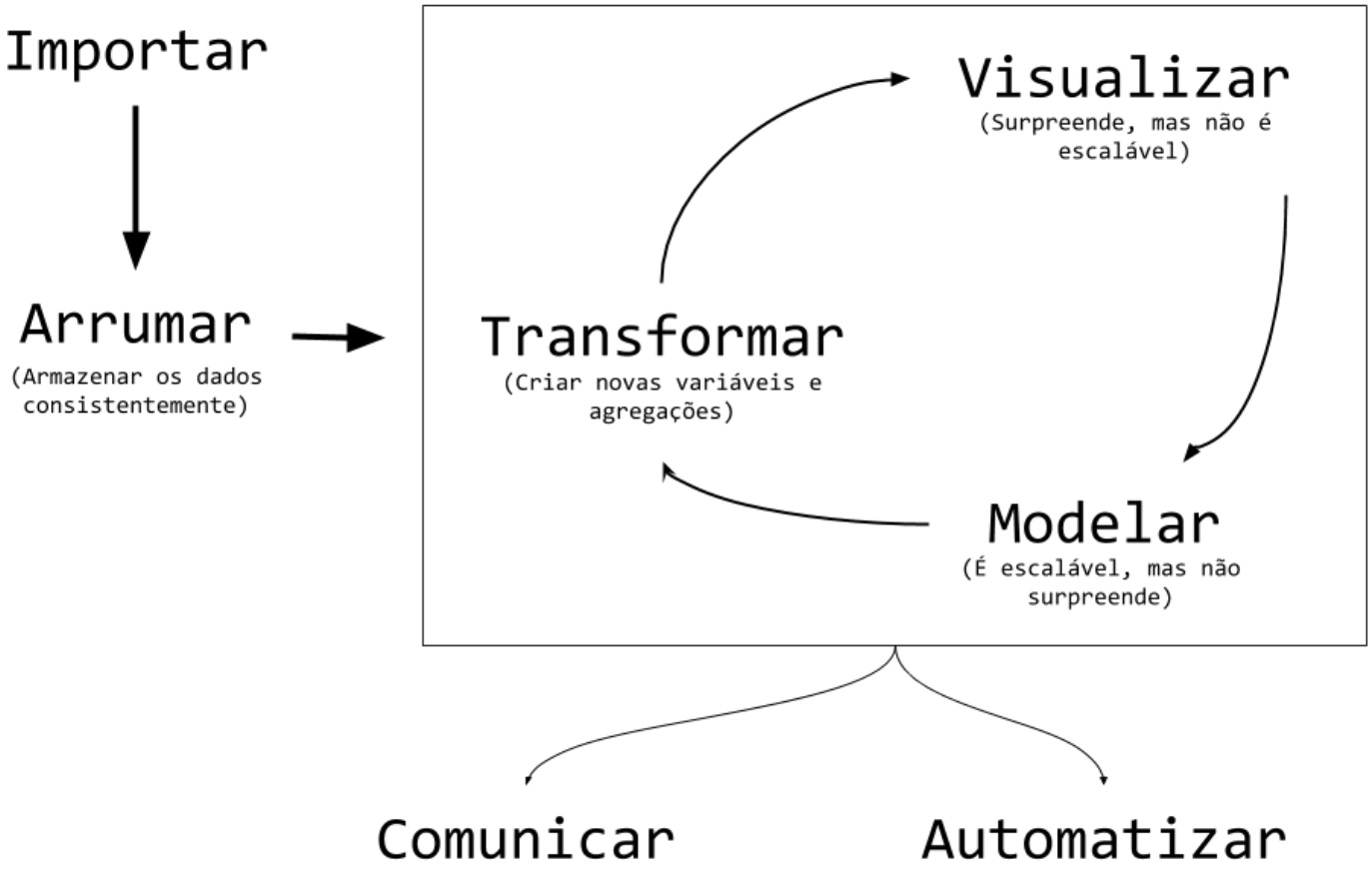
Objetivos: Visualização e modelagem de dados a partir de uma base de uma lista de processos enviada pela Vivo.

Recorte temporal: processos distribuídos entre 2009 e 2019.

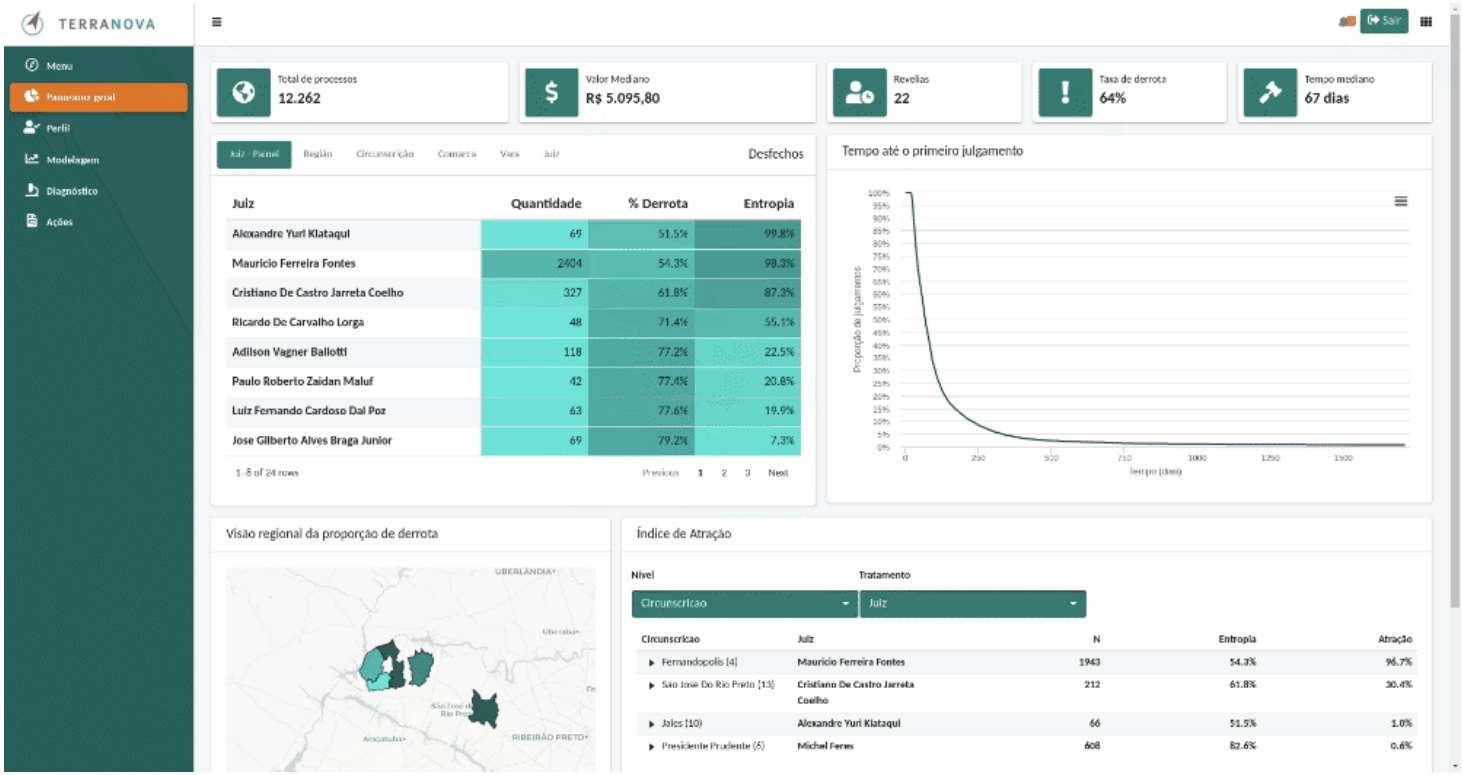
Recorte regional: Processos distribuídos em 6 comarcas distintas: Fernandópolis, Jales, Palmeira D'Oeste, Presidente Prudente, Santa Fé do Sul e São José do Rio Preto.

Recorte de escopo: Processos relacionados a direito do consumidor com a Vivo no polo passivo.

Ciclo da ciência de dados



O Dashboard


TERRANOVA
Sair
Menu
Panorama geral
Perfil
Modelagem
Diagnóstico
Ações
Total de processos
12.262
Valor Mediano
R$ 5.095,80
Revelias
22
Taxa de derrota
64%
Tempo mediano
67 dias
Juiz - Painel
Região
Circunscrição
Comarca
Vara
Juiz
Desfechos
Juiz
Quantidade
% Derrota
Entropia
Alexandre Yurl Klataqui 69 51.5% 99.8%
Mauricio Ferreira Fontes 2404 54.3% 98.3%
Cristiano De Castro Jarreta Coelho 327 61.8% 87.3%
Ricardo De Carvalho Lorga 48 71.4% 55.1%
Adilson Vagner Ballotti 118 77.2% 22.5%
Paulo Roberto Zaidan Maluf 42 77.4% 20.8%
Luiz Fernando Cardoso Dal Poz 63 77.6% 19.9%
Jose Gilberto Alves Braga Junior 69 79.2% 7.3%
1-8 of 24 rows
Previous 1 2 3 Next
Tempo até o primeiro julgamento
Proporção de julgamentos
Tempo (dias)
Visão regional da proporção de derrota
UBERLÂNDIA
RIBEIRÃO PRETO
Índice de Atração
Nível
Tratamento
Circunscricao
Juiz
Circunscricao
Juiz
N
Entropia
Atração
Fernandopolis (4) Mauricio Ferreira Fontes 1943 54.3% 96.7%
Sao Jose Do Rio Preto (13) Cristiano De Castro Jarreta Coelho 212 61.8% 30.4%
Jales (10) Alexandre Yurl Klataqui 66 51.5% 1.0%
Presidente Prudente (6) Michel Feres 608 82.6% 0.6%

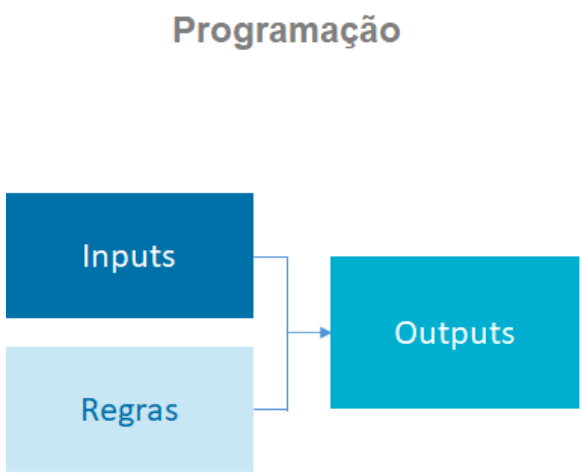
Modelagem preditiva

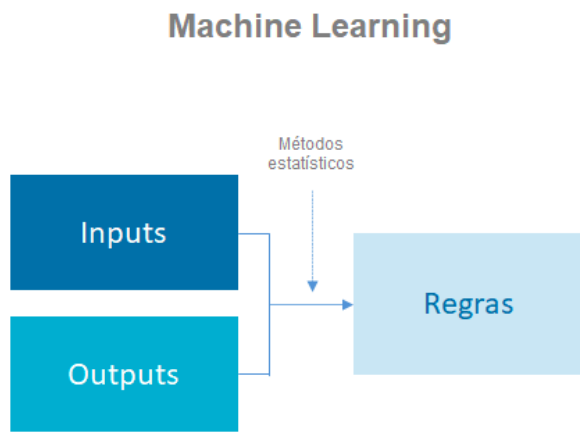
- Área da estatística destinada à construção de **modelos estatísticos** capazes de fornecer boas predições para determinado fenômeno.
- Melhor introdução para o tema: ISLR

- **Machine learning** e **statistical learning** são **sinônimos**.

O que é um modelo preditivo?

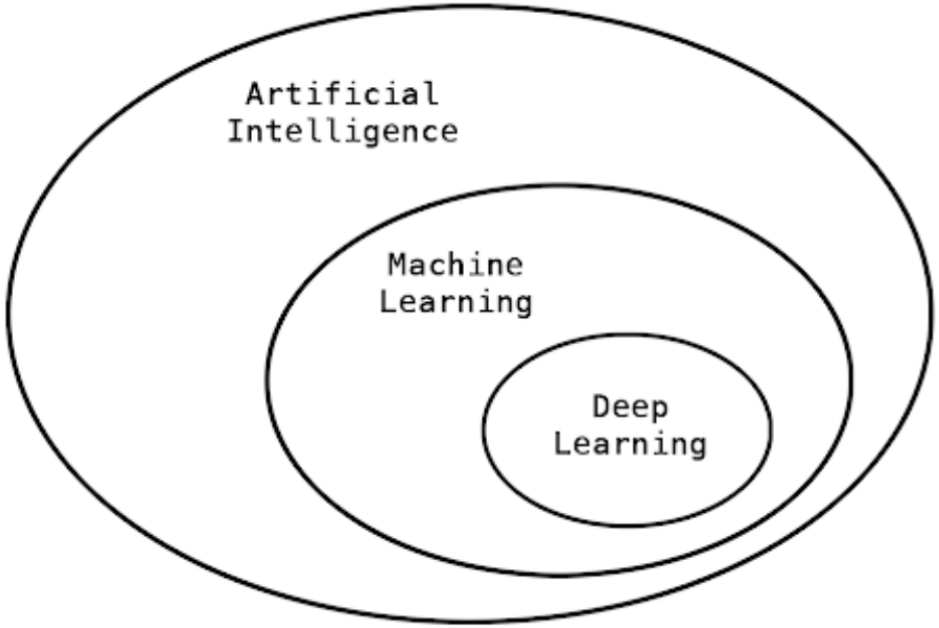
Paradigma programação vs paradigma do aprendizado estatístico.



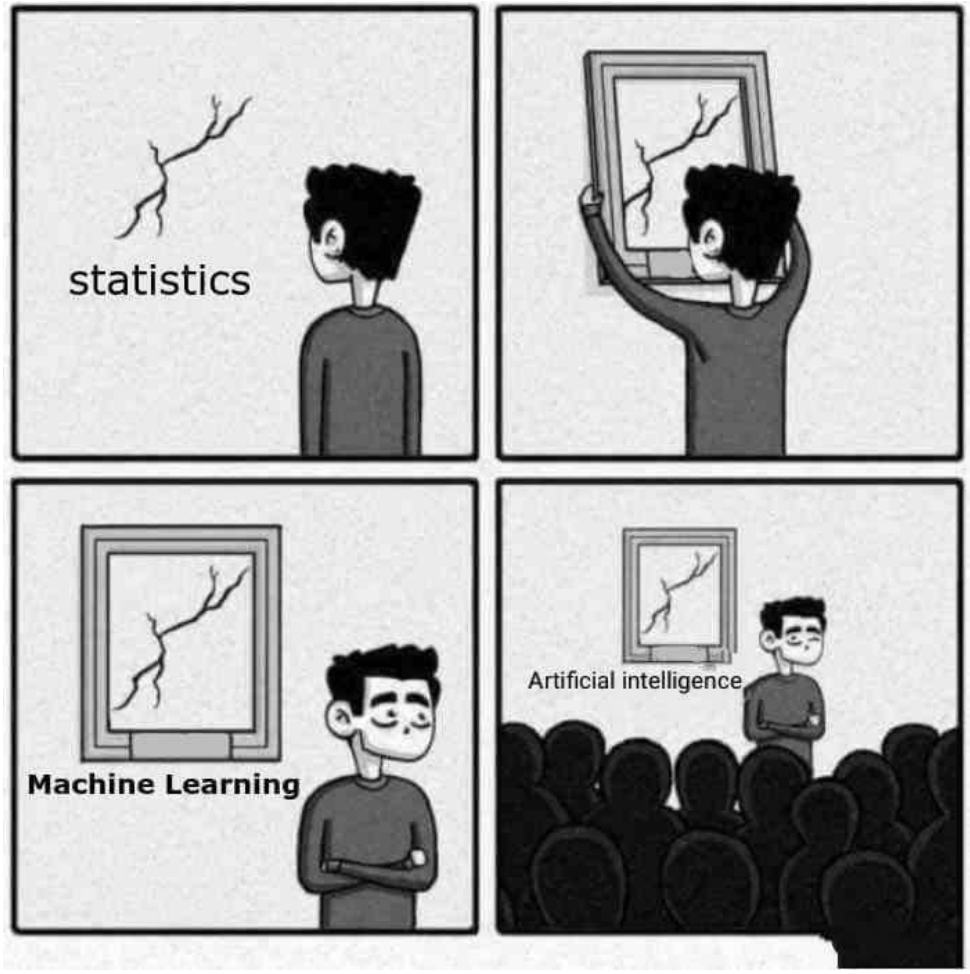


Mas e as redes neurais?

- Redes neurais / deep learning é uma área de aprendizado de máquinas (aprendizado estatístico).



Mas e a inteligência artificial?!



A inteligência artificial *está* machine learning nos dias de hoje.

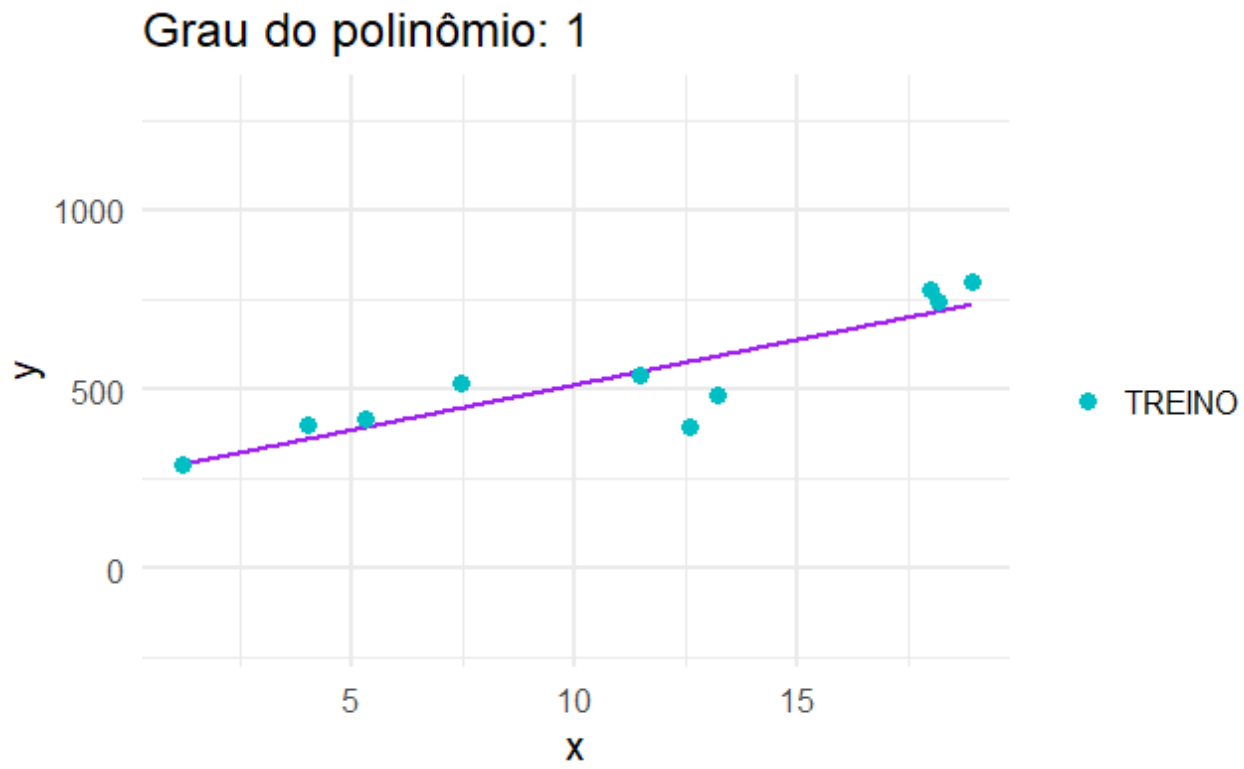
A única fórmula que você verá aqui

- Nós temos inputs e outputs, e queremos predizer um output que ainda não existe a patir de um input que temos em mãos.
- Por exemplo, queremos predizer o resultado de um processo (output) com base nas suas características (input).
- Geralmente associamos os inputs à variável X e os outputs à variável Y.
- Queremos criar uma função (uma fórmula) g que, a partir de um X, consegue dar um valor de Y

$$Y = g(X)$$

Sobreajuste

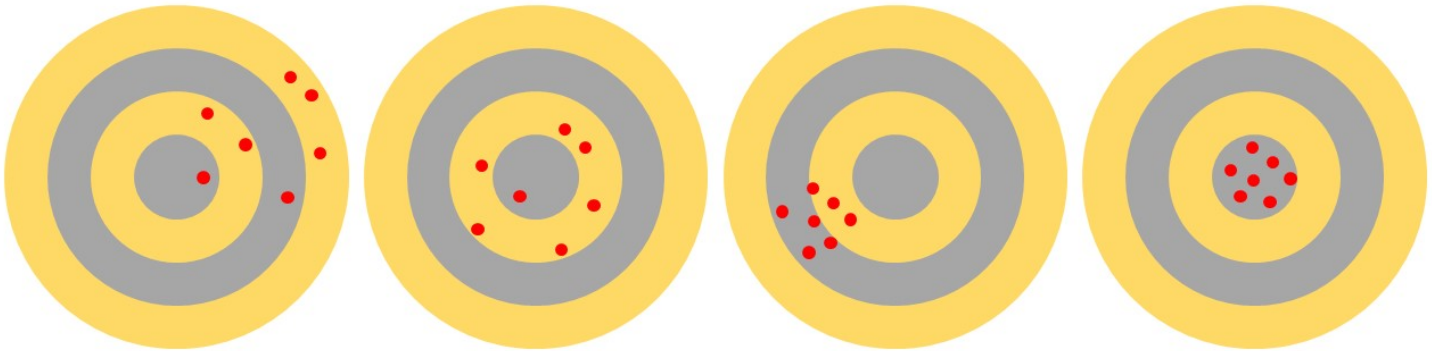
Para isso, podemos aplicar um modelo super complexo, que se ajusta perfeitamente aos dados que eu observo, como se cada caso fosse um caso...



Fonte: Curso-R

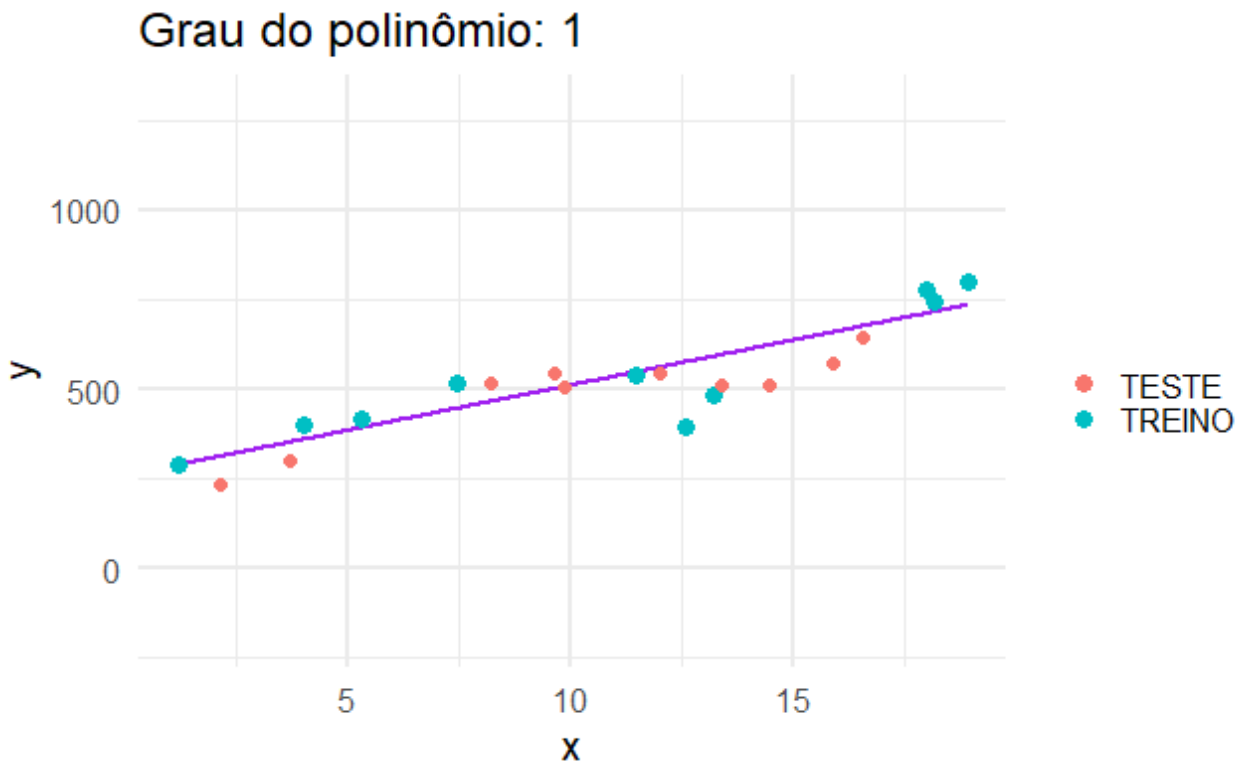
Erro de generalização

- Um modelo de aprendizado estatístico precisa funcionar bem para bases que nós **não observamos**. Para isso, tentamos criar um modelo que se adeque bem aos dados que observamos.



Sobreajuste

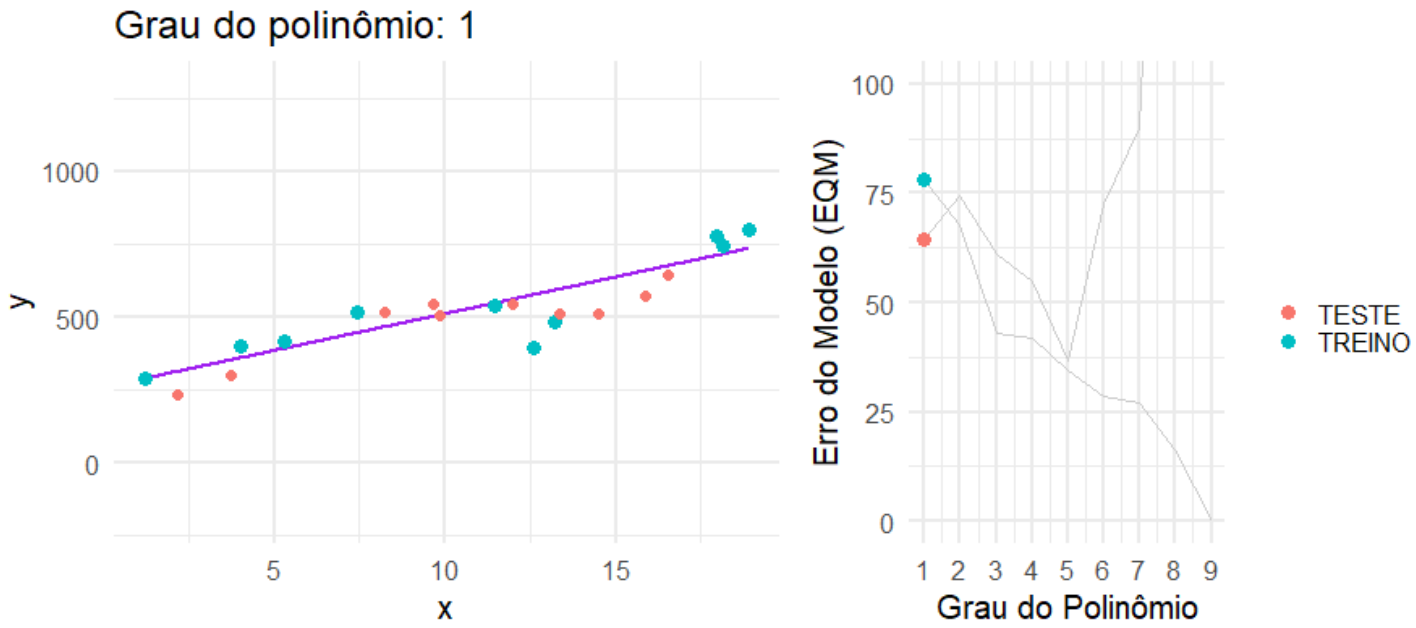
... mas quando eu vou aplicar isso no mundo real, os modelos mais complicados não se aplicam.



Fonte: Curso-R

Sobreajuste

No fundo, preciso escolher um modelo que seja suficientemente complexo para captar o **sinal** (tendência) do fenômeno estudado, sem com isso fazer com que meu modelo seja suscetível a **ruídos** (erros aleatórios),



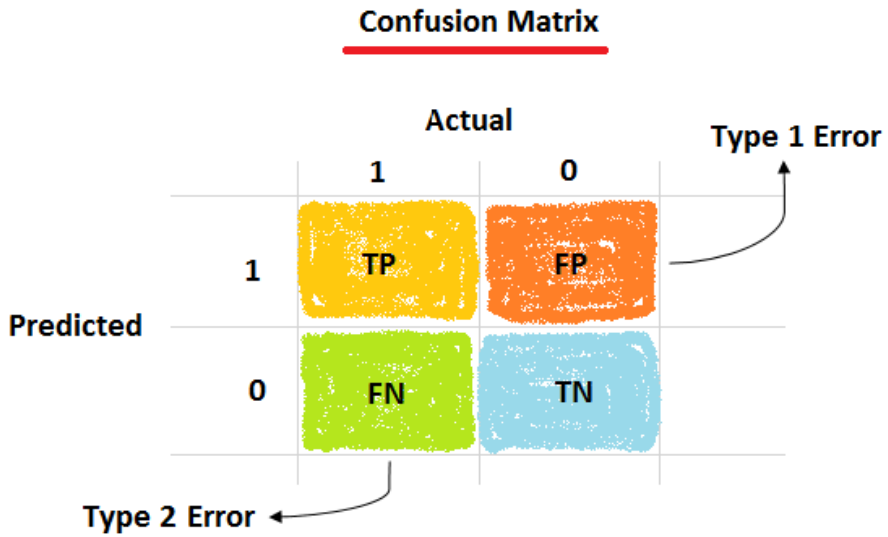
Fonte: Curso-R

Treino e teste

- Para lidar com esse problema, separamos nossa base em duas: uma base de **treino** e uma base de **teste**.
- Na base de treino, ajustamos nosso modelo. Na base de teste, testamos o quão bom está o modelo. Dessa forma, não tem como sermos enganados pelo sobreajuste.
- A base de treino, por sua vez, passa por um procedimento chamado **validação cruzada**, que consiste em testar vários modelos escondendo uma parte da base de treino e testando os melhores candidatos.

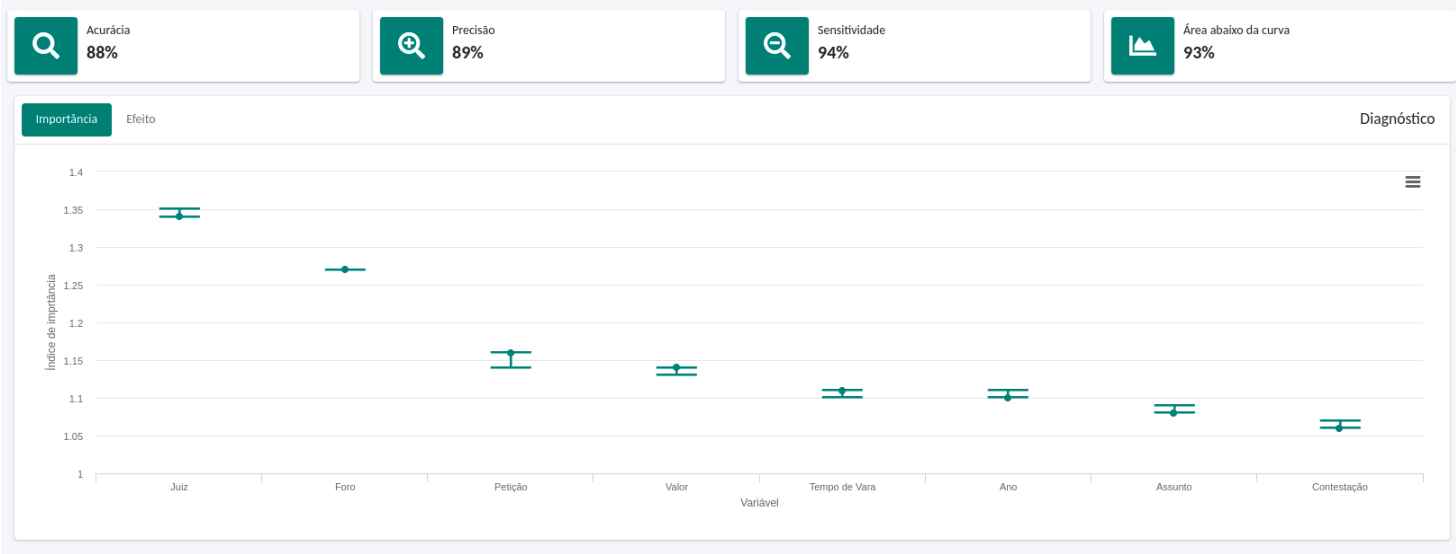
Métricas de qualidade de um modelo

- Nem sempre o melhor modelo é aquele que acerta mais!
- A métrica depende do problema estatístico (regressão, classificação) e do **problema de negócio**.
- **Exemplo**: acurácia, falso positivo e falso negativo



Fonte: Machine learning Model Performace Metrics

Voltando ao case



- Acurácia do modelo: 88%. É bom?
- Precisão do modelo: 89%.

Aprendizado estatístico interpretável

Art. 20 da LGPD:

> § 1º O controlador deverá fornecer, sempre que solicitadas, informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada, observados os segredos comercial e industrial.

- Conforme obtemos mais dados, conseguimos ajustar modelos cada vez mais complexos.
- Modelos mais complexos são mais difíceis de interpretar: não tem só um X para explicar o Y, e as funções g podem ser bem complicadas de ler por seres humanos.

Machine Leaning Interpretável

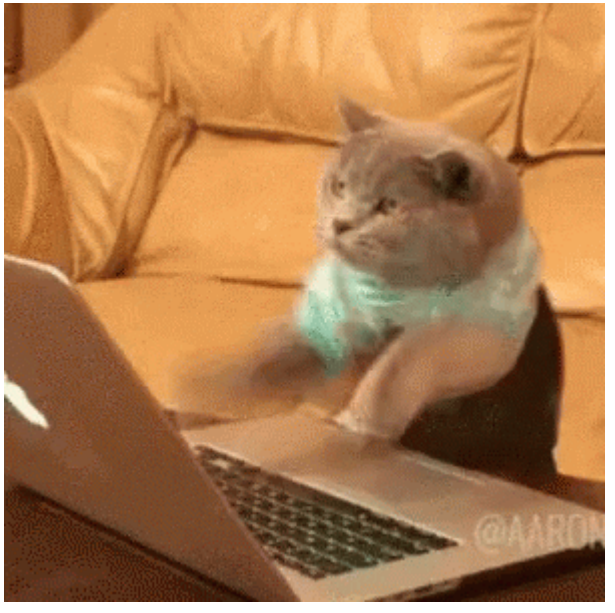
Uma das áreas de pesquisa que mais cresce na atualidade é a de intepretação de modelos. O problema de *fairness* é igualmente relevante, e também é levada em conta na área de interpretabilidade.

Livro: Interpretable Machine Learning.

Resumo

- Aprendizado de máquinas é o mesmo que **aprendizado estatístico**, e inteligência artificial é uma área do conhecimento que contém como área mais importante (atualmente) o aprendizado de máquinas.
- Modelos preditivos buscam minimizar o erro de generalização. Para isso, precisamos separar a base entre **treino** e **teste**.
- Um modelo pode otimizar **métricas diferentes**, e essas métricas dependem do modelo de negócio.

Quiz



https://forms.office.com/r/dYnft5E4Nf

Obrigado!

Julio Trecenti