



CIÊNCIA DE DADOS APLICADA A ANÁLISE ESPORTIVA
UTILIZANDO PYTHON AVANÇADO

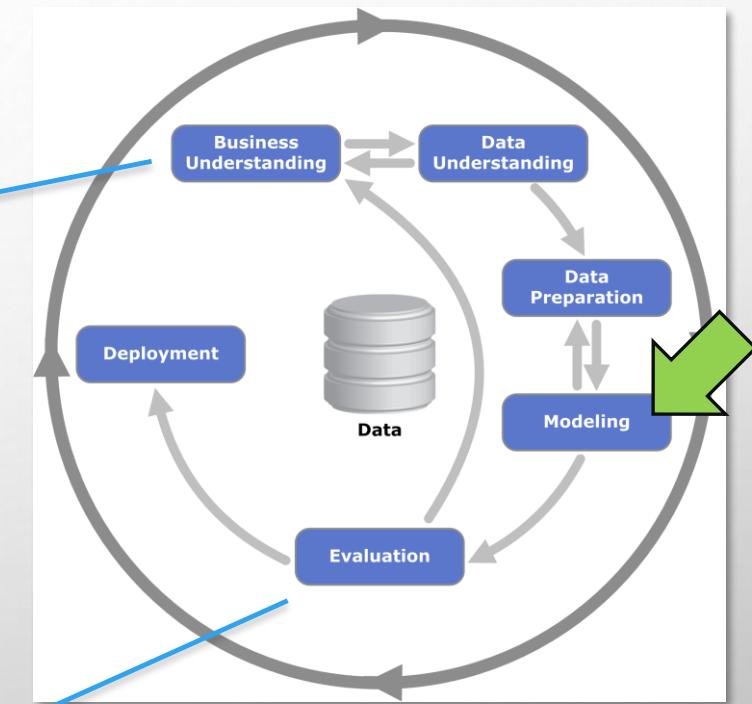
INTRODUÇÃO AO APRENDIZADO DE MÁQUINA

DIEGO RODRIGUES DSC

INFNET

CRONOGRAMA

NÚMERO	ÁREA	AULA	TRABALHOS
1	Estatística	Intro	Introdução a Disciplina e Organização do Ambiente
2		Dados	Coleta de Dados e Sensoriamento
3			Variáveis Aleatórias
4			Análise Exploratória
5			Estatísticas para Ranqueamento
6			Rankeamento Estatístico : ELO
7			Rankeamento Estatístico : Glicko
8			Rankeamento Estatístico : TrueSkill
9			Rankeamento Estatístico : XELO
10	ML		Base de Dados
11			Modelos de Aprendizado de Máquina
12			Machine Learning: Classificação
13			Machine Learning: Regressão
14			Machine Learning: Agrupamento
15	Esportes		Pesquisa
16			Machine Learning: Visão Computacional
17			Aplicações & Artigos: Esportes Independentes
18			Aplicações & Artigos: Esportes de Objeto
19			Aplicações & Artigos: Esportes de Combate
			Aplicações & Artigos : Betting
			Workshop



CRISP-DM

AMBIENTE PYTHON



4. Variáveis Aleatórias



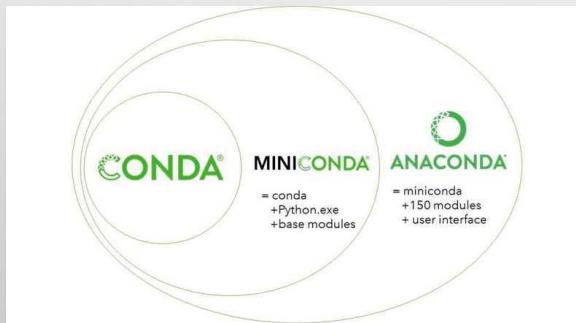
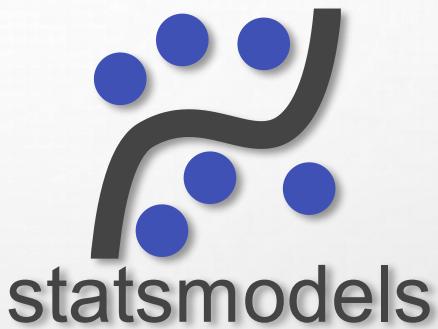
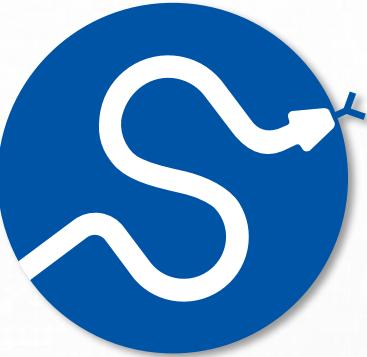
1. Editor de Código



5. Visualização



6. Estimação e Inferência



2. Gestor de Ambiente



3. Ambiente Python do Projeto



7. Machine Learning

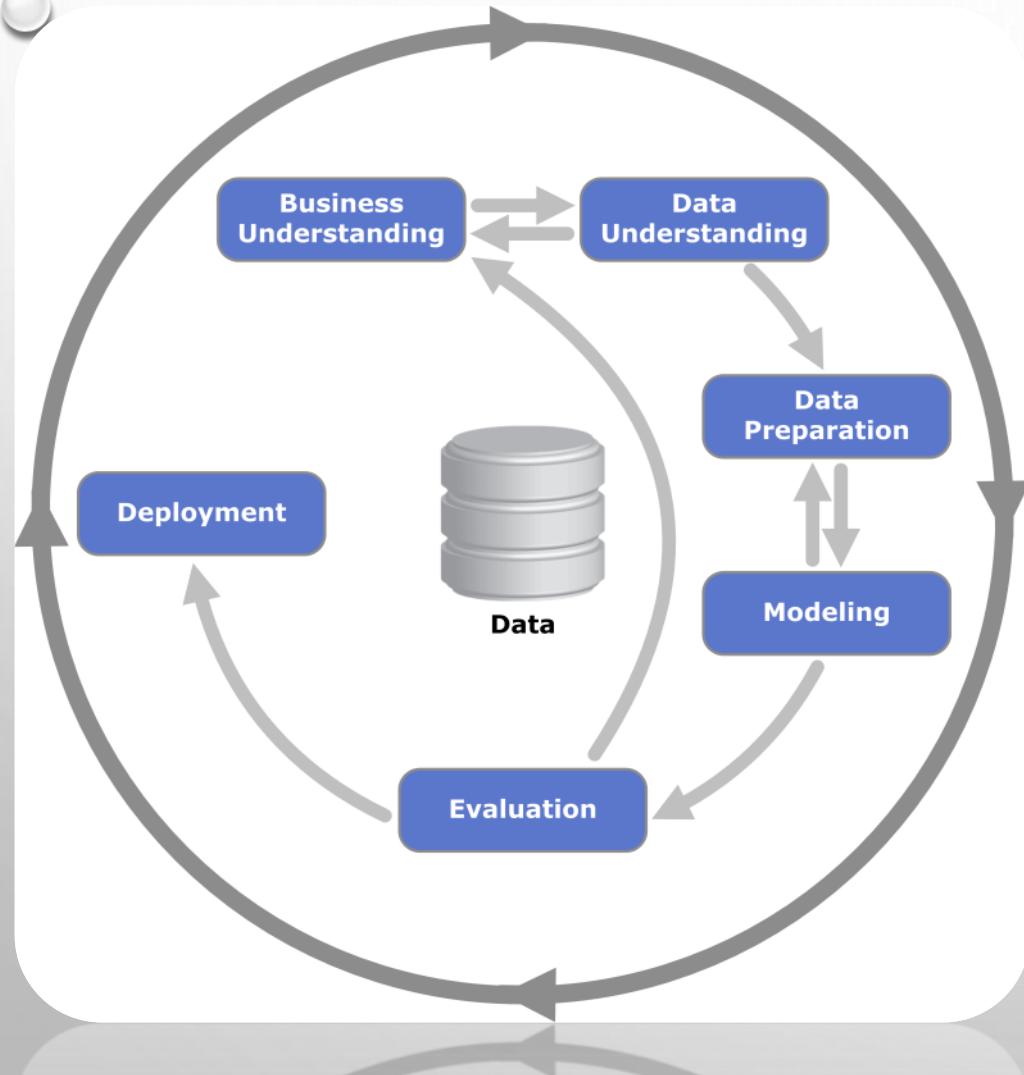


3. Notebook Dinâmico

PROCESSO DE DESENVOLVIMENTO

CRISP-DM

Cross Industry Standard Process for Data Mining - IBM



1) Requerimentos e Análise de Negócio

Entendimento do problema decisório, dados relacionados & revisão bibliográfica.

2) Preparação dos Dados

Entendimento das fontes de dados, dos tipos, análise exploratória e representação.

3) Modelagem

Seleção, extração de atributos e treinamento do modelo.

4) Avaliação

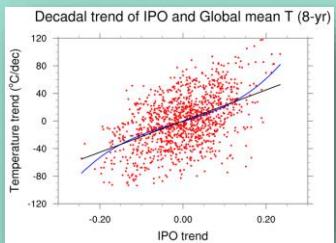
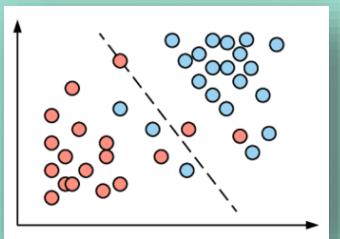
Seleção do melhor modelo.

5) Liberação

Liberação do modelo no ambiente de produção.

BUSINESS UNDERSTANDING

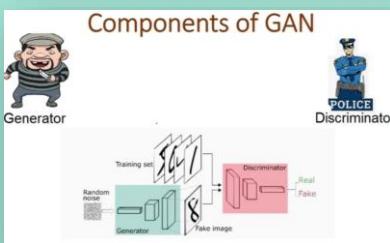
APRENDIZADO SUPERVISIONADO



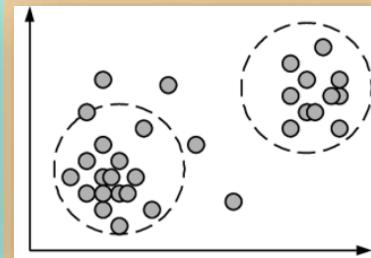
CLASSIFICAÇÃO

REGRESSÃO

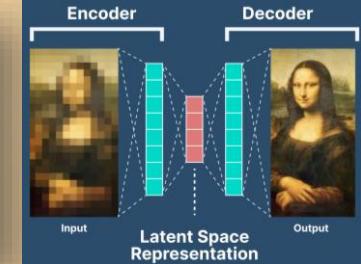
GENERATIVO



APRENDIZADO NÃO-SUPERVISIONADO

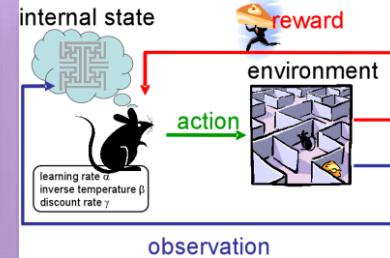


AGRUPAMENTO



ENCODER
DECODER

APRENDIZADO POR REFORÇO



REFORÇO

APRENDIZADO SUPERVISIONADO

Tarefas de **classificação** e **regressão** pertencem a esta categoria.

O treinamento consiste em **encontrar parâmetros** para o modelo que **minimiza uma função de risco/erro** para uma amostra de treinamento, baseado na diferença entre os **valores previstos e reais**, para cada observação.

CLASSIFICAÇÃO

Um bebê consegue separar e ordenar blocos com diferentes tamanhos, formas e cores. Ele também consegue **identificar os tipos diferentes de objetos.**

Os diferentes tipos de objetos são chamados de **classes**. As características dos objetos são chamadas de **variáveis**, ou **atributos**.

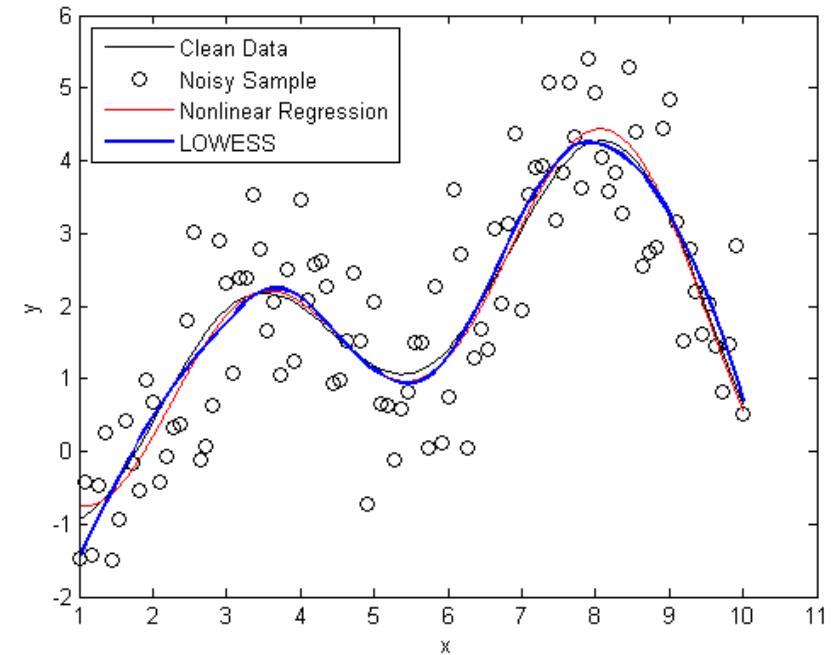


Então, um classificador é um modelo **treinado para discriminar objetos** pertencentes a duas ou mais classes, baseado em seus atributos.

REGRESSÃO

O objetivo da regressão é
modelar as relações funcionais
entre dois conjuntos de variáveis.

As variáveis que representam as causas são
chamadas de **variáveis independentes**, e as
variáveis cujo objetivo é prever, são chamadas
variáveis dependentes.

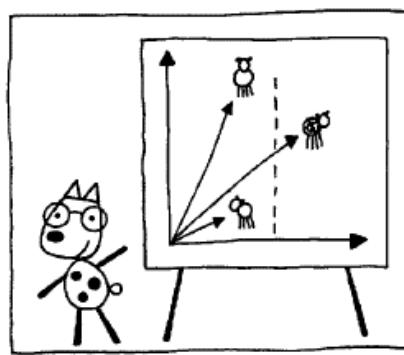
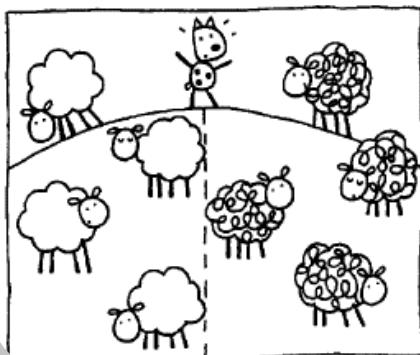
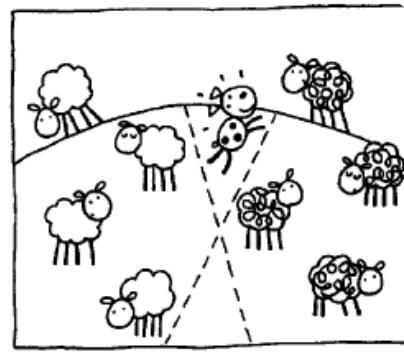
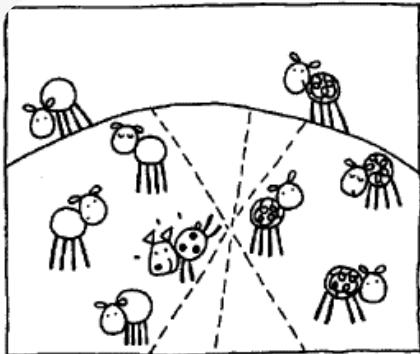


As vezes quando o mundo
não é linear & gaussiano...

Então, uma **regressão** é um modelo utilizado para
prever **uma ou mais variáveis dependentes**,
baseado em causas, ou variáveis independentes.

DATA PREPARATION

REPRESENTAÇÃO



Ideia: como quantificar um objeto
no mundo físico no mundo digital?

Exercício: qual seria uma boa representação
para diferenciar ratos e elefantes?

DATA PREPARATION

Quantificação dos Atributos

- Transformar todos os atributos em atributos numéricos.

Normalização

- Transformar todos os atributos para a mesma faixa dinâmica, de maneira a assegurar que todos tenham o mesmo “peso numérico” para o treinamento do modelo.

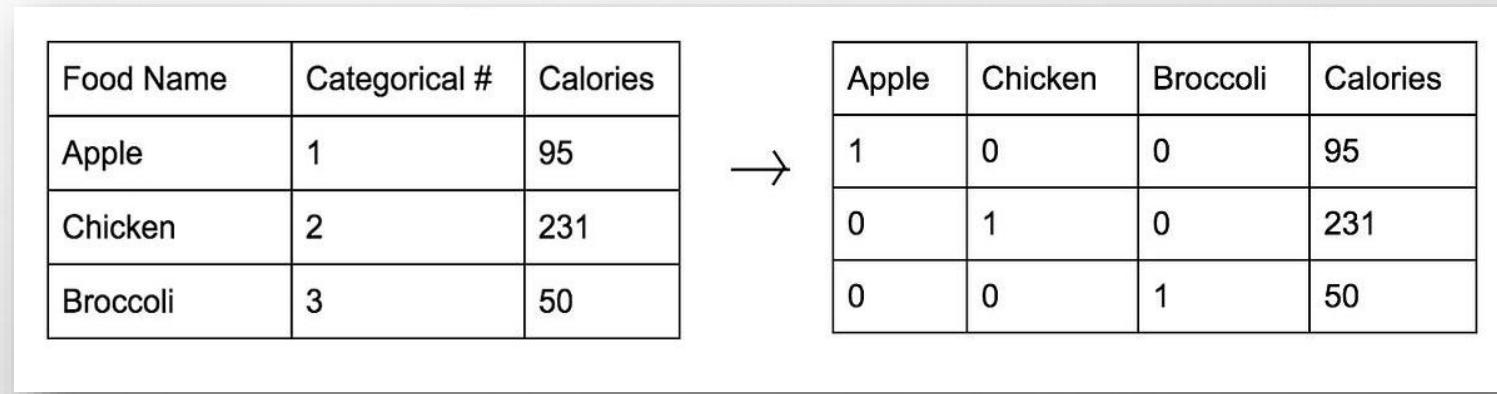
ATRIBUTOS CATEGÓRICOS

One Hot Encoding

Gender
Female
Male
Male
Female



Gender
1
0
0
1



DATAS

Componentes da Data

- Ano
- Mês
- Dia
- Dia do Ano
- Dia da Semana
- Hora
- Minuto
- Segundo

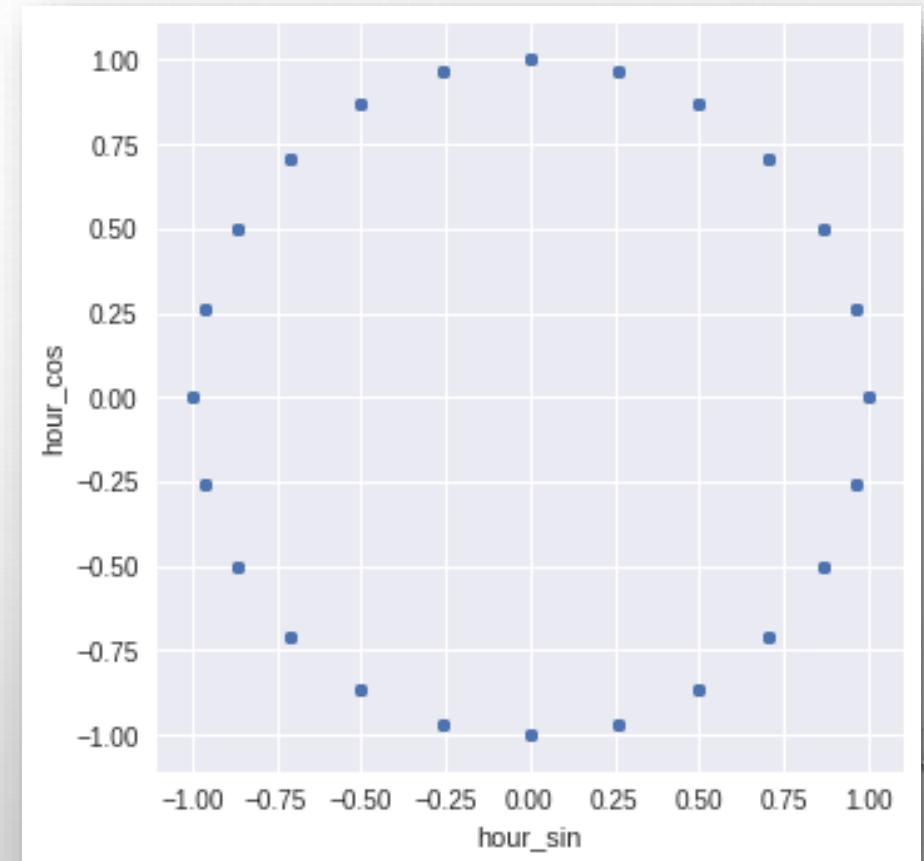
Flags

- É final de semana
- É feriado

Diferença entre Datas

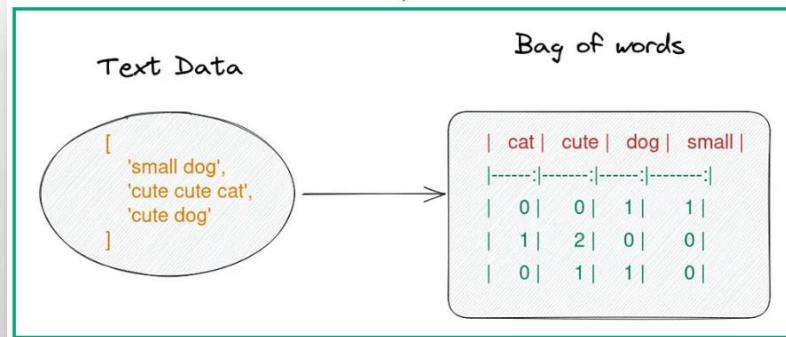
- Diferença em Dias
- Diferença em Horas
- Diferença em Meses

Encoding Cíclico



ATRIBUTOS TEXTUAIS

BAG OF WORDS



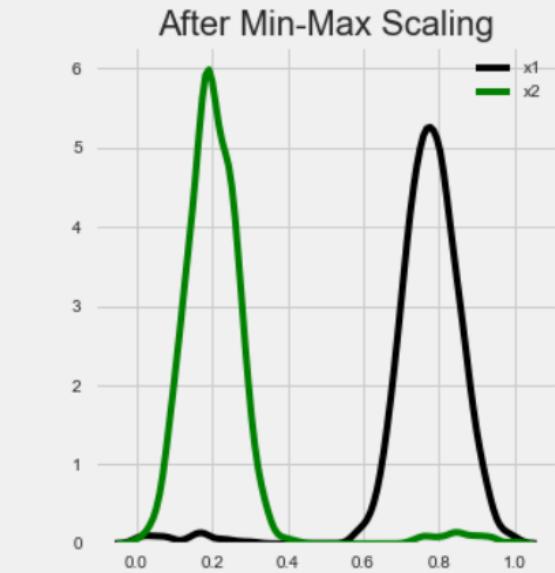
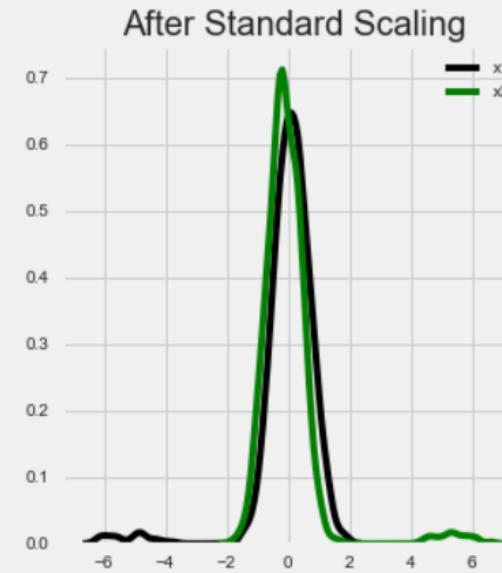
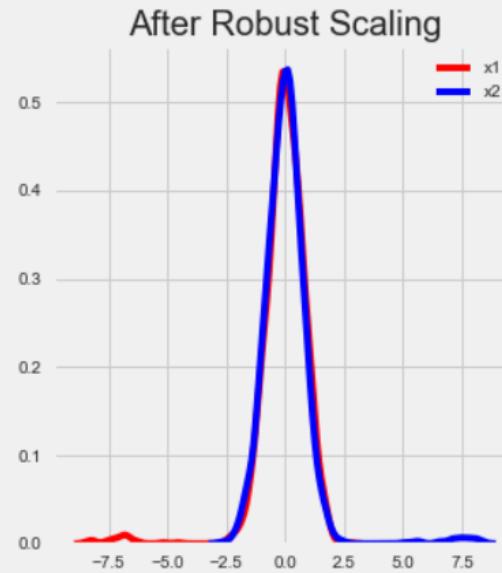
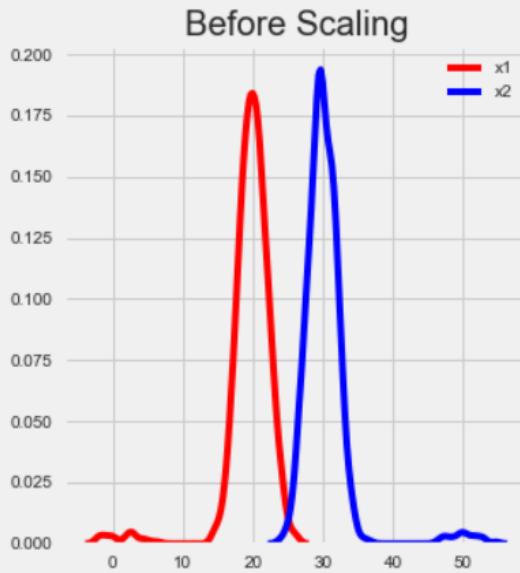
Variants of term frequency (tf) weight	
weighting scheme	tf weight
binary	0, 1
raw count	$f_{t,d}$
term frequency	$f_{t,d} / \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Variants of inverse document frequency (idf) weight	
weighting scheme	idf weight ($n_t = \{d \in D : t \in d\} $)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log \left(\frac{N}{1 + n_t} \right) + 1$
inverse document frequency max	$\log \left(\frac{\max_{\{t' \in d\}} n_{t'}}{1 + n_t} \right)$
probabilistic inverse document frequency	$\log \frac{N - n_t}{n_t}$

Words	Count		Term Frequency (TF)		Inverse Document Frequency (IDF)	TF * IDF	
	Document 1	Document 2	Document 1	Document 2		Document 1	Document 2
read	1	1	0.17	0.17	0	0	0
svm	1	0	0.17	0	0.3	0	0.05
algorithm	1	1	0.17	0.17	0	0	0
article	1	1	0.17	0.17	0	0	0
dataaspirant	1	1	0.17	0.17	0	0	0
blog	1	1	0.17	0.17	0	0	0
randomforest	0	1	0	0.17	0.3	0	0.05

NORMALIZAÇÃO

- Garantir que as variáveis possuam a mesma escala
- Mesmo efeito numérico na otimização independente da escala.
- Transformar de outra distribuição para distribuição normal



TÉCNICAS DE SELEÇÃO DE ATRIBUTOS

Filtragem – mede a relação entre atributos ou atributos e classes, utilizando estatísticas, sem depender do modelo.

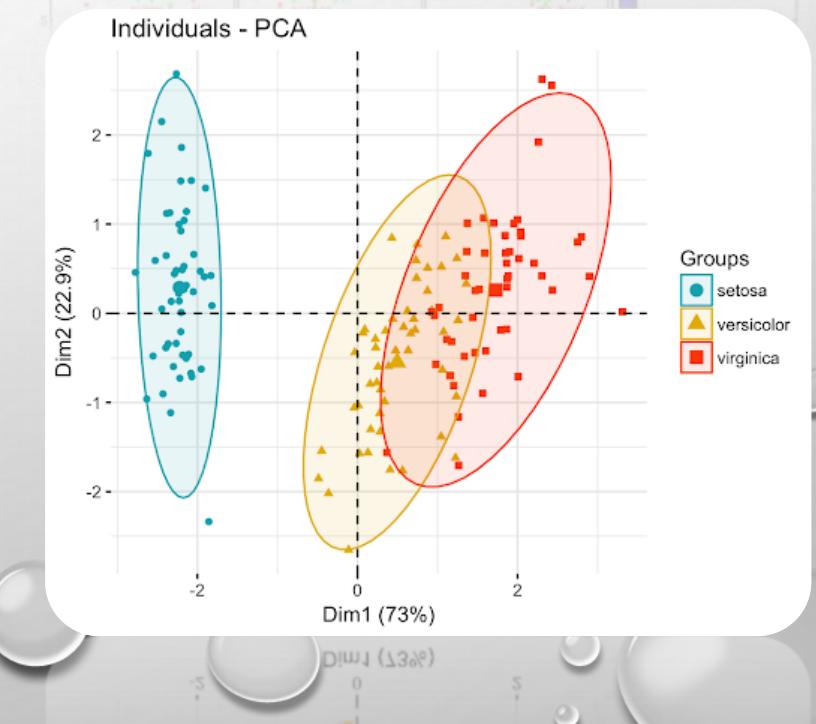
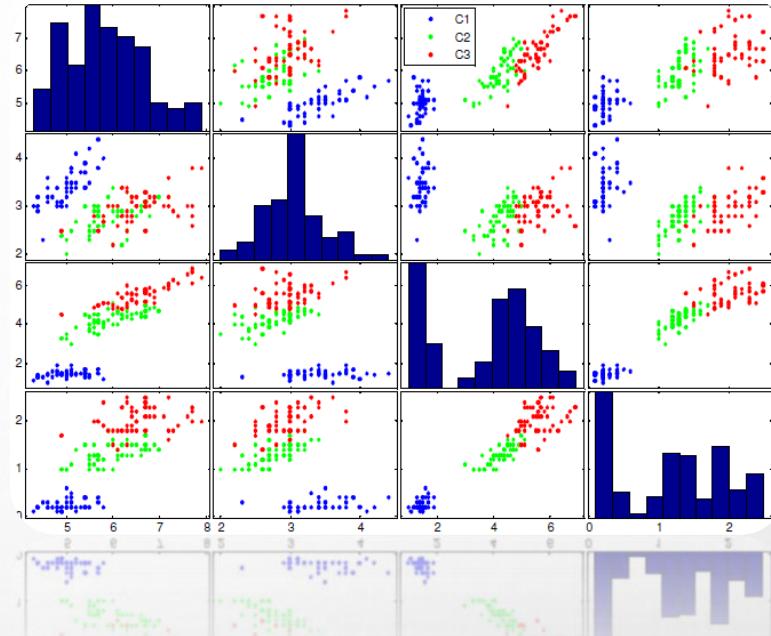
- **Coeficiente de Correlação de Pearson** – Estatística que mede a relação linear entre duas variáveis aleatórias.
- **Teste T de diferença de médias** – Informa se a média de um determinado atributo muda de acordo com uma categoria binária.
- **ANOVA** – O mesmo que o teste T, mas serve para múltiplas categoria.
- **Informação Mútua** – Estatística que mede relação não-linear entre duas variáveis aleatórias.

Wrapper – mede a relação entre atributos e classes, utilizando um modelo treinado.

- **Gini** – Estatística que representa a importância de um atributo na divisão da base de dados por uma árvore de decisão.
- **Relevância** – Estatística que representa a variação causada na saída do modelo quando um atributo é substituído por sua média.

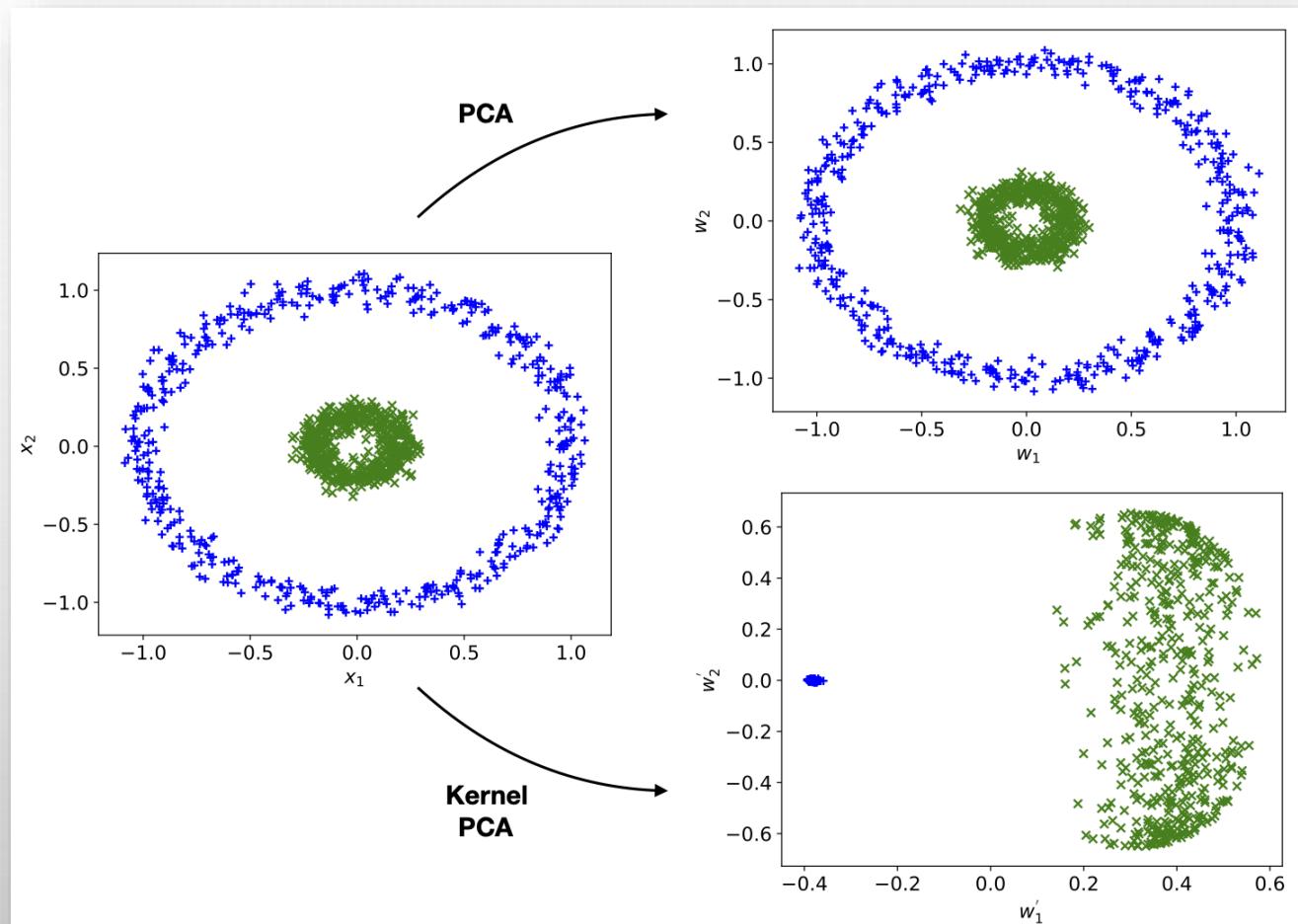
EXTRAÇÃO DE ATRIBUTOS ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

- Garantir que as variáveis independentes sejam descorrelacionadas.
- Identificar novas direções com maior concentração de energia / informação.
- Variáveis transformadas perdem o sentido físico.



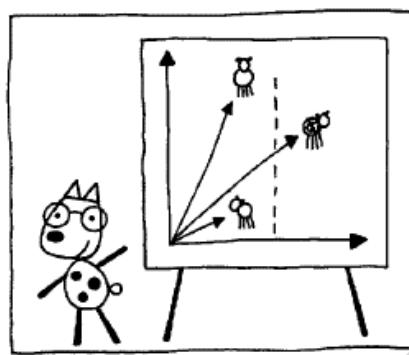
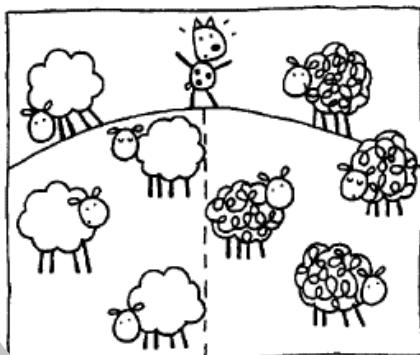
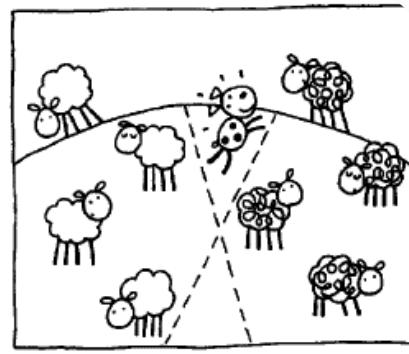
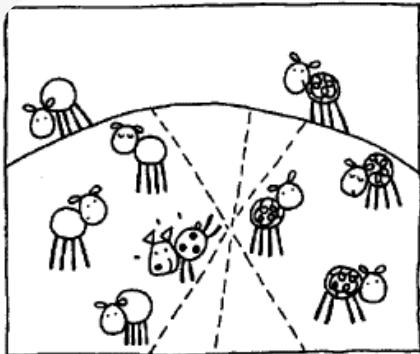
EXTRAÇÃO DE ATRIBUTOS – KERNEL PCA

- Identifica novo espaço que favoreça a modelagem.
- Como selecionar o Kernel Adequado?



CLASSIFICAÇÃO

EXERCÍCIO



Exercício: qual seria uma boa representação para diferenciar ratos e elefantes?

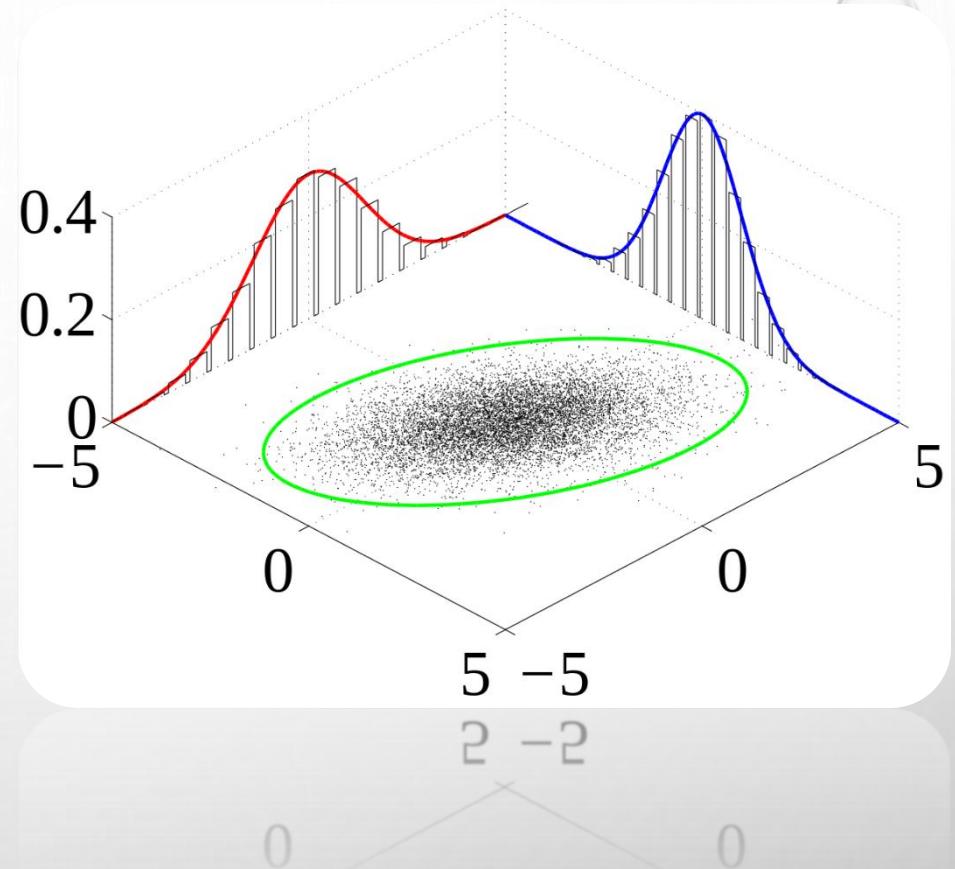
**Visualização de Algoritmos em
Ratos vs Elefantes**

MODELING

ALGORITMOS BASEADOS EM DENSIDADE

Algoritmos que dependem da **função densidade de probabilidade** dos dados, ou aproximações locais, para determinar a classe de observações fora da amostra de treino.

- 1) Classificador Bayesiano
- 2) Classificador Bayesiano “Naïve”
- 3) K-Vizinhos mais próximos

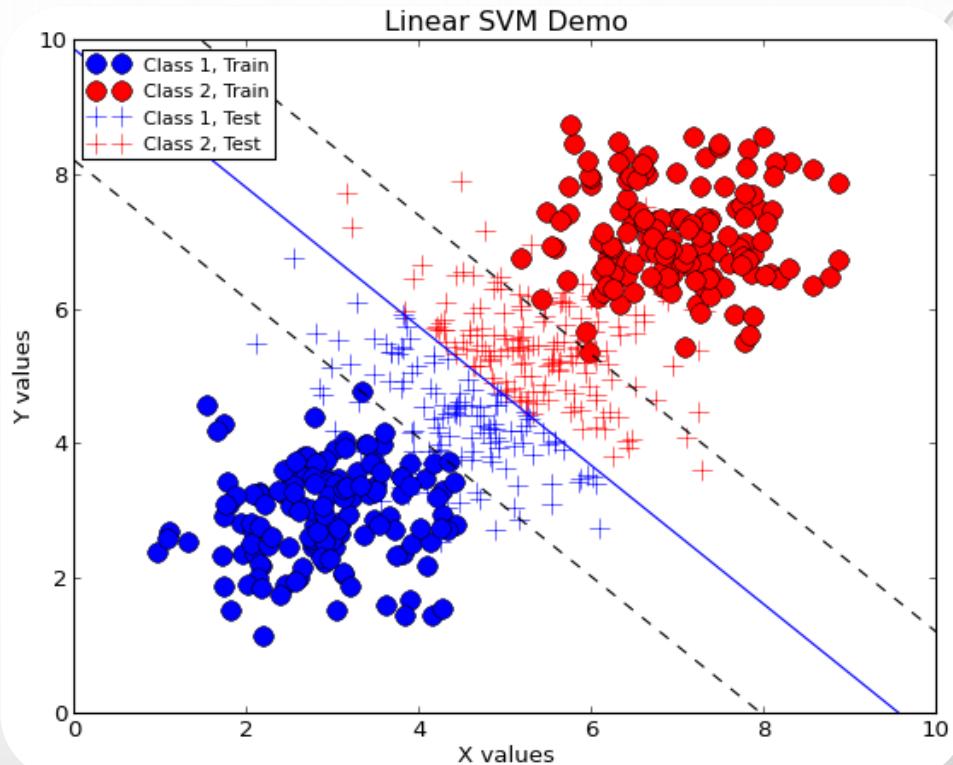


Algoritmos baseados em densidade dependem da **DENSIDADE** (!!). Consequentemente, se beneficiam de um **conjunto grande de observações e de baixa esparsidade do espaço de atributos**. O Classificador Bayesiano é considerado o classificador “ótimo”, mas é raramente utilizado, dada a dificuldade de estimar a função densidade de probabilidade dos dados. É normalmente utilizado como benchmark para comparação teórica entre os algoritmos de classificação.

MODELOS FUNCIONAIS

Algoritmos que dependem da **estimação dos parâmetros de uma função** que é utilizada como **superfície de separação** entre as classes.

- 1) Funções Polinomiais
- 2) **Regressão Logística**
- 3) Máquina de Vetores Suporte
- 4) **Neurônio Sigmoide / Tangente Hiperbólica**
- 5) **Árvores de Decisão**



Algoritmos baseados em funções são **mais simples**, usualmente tem um **número menor de parâmetros** e não dependem em armazenar muitos dados para manter uma “memória”, como por exemplo K-vizinhos mais próximos.

ALGORITMOS BASEADOS EM ENSEMBLE

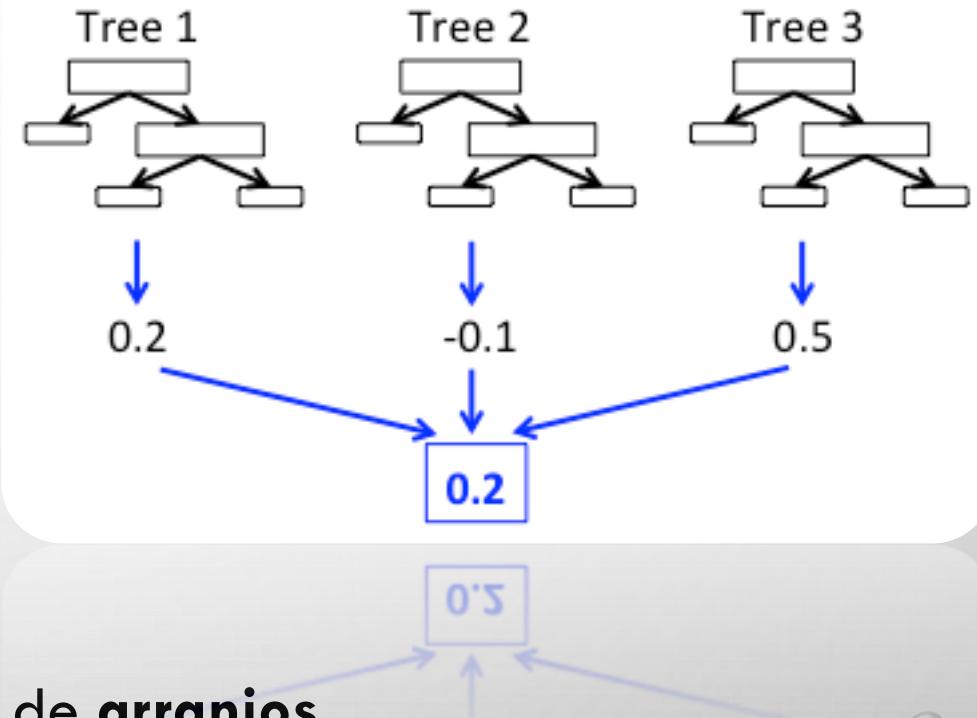
Algoritmos que **combinam modelos simples**,
usualmente através de **votação ou ponderação**, para
atingir maiores taxas de classificação.

1) Random Forest

2) Boosting

Boa **capacidade de generalização** gerado através de **arranjos complexos** de múltiplos modelos simples de machine learning.

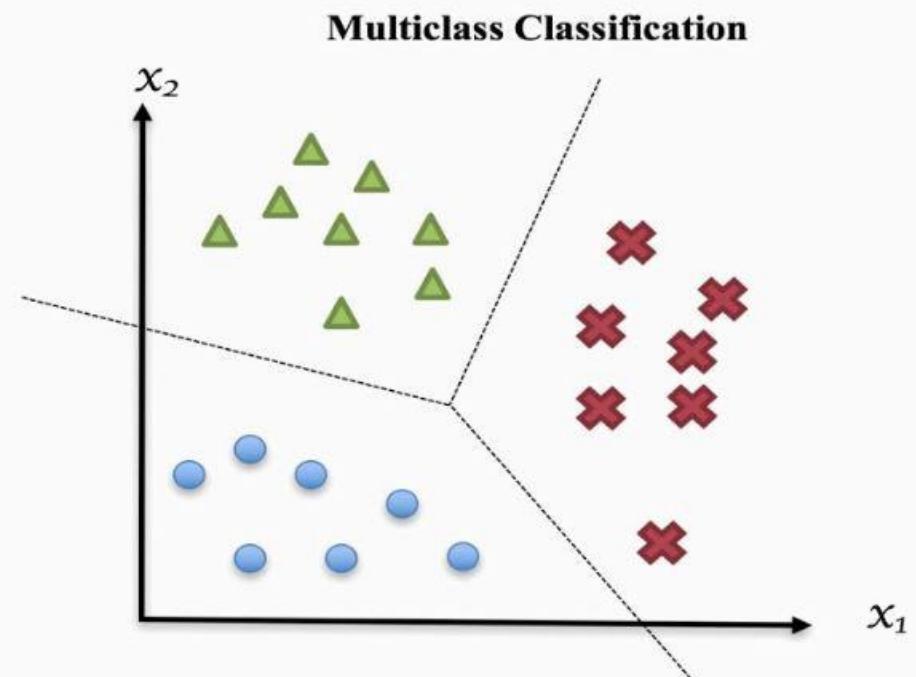
Ensemble Model:
example for regression



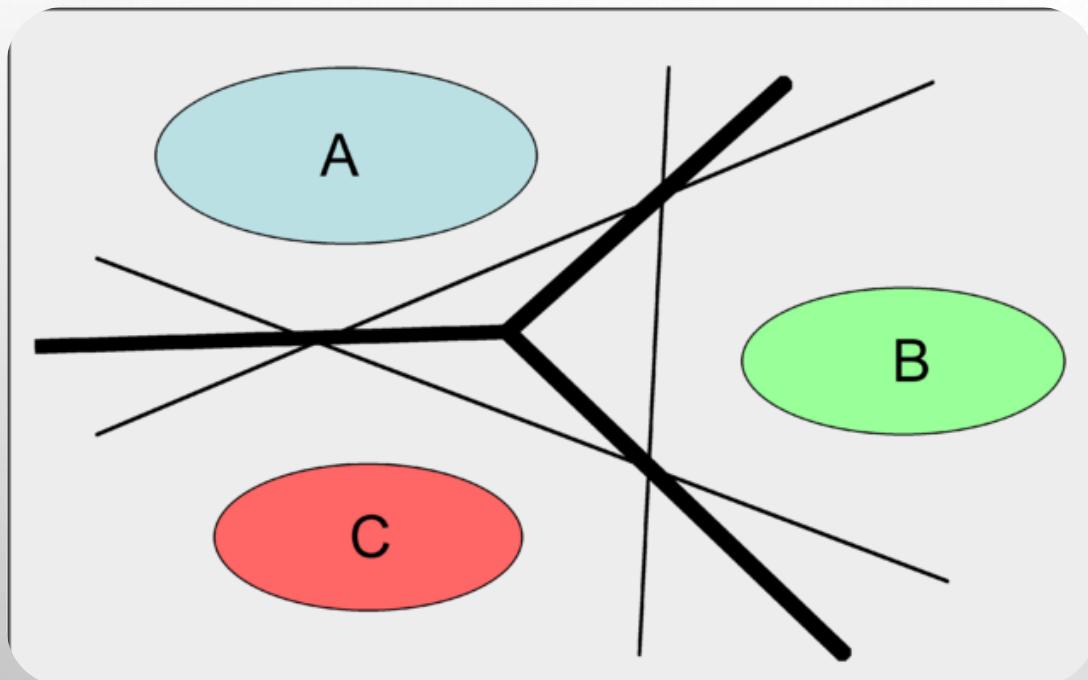
ALGORITMOS BASEADOS EM ENSEMBLE

- **Modelos Multiclasse**

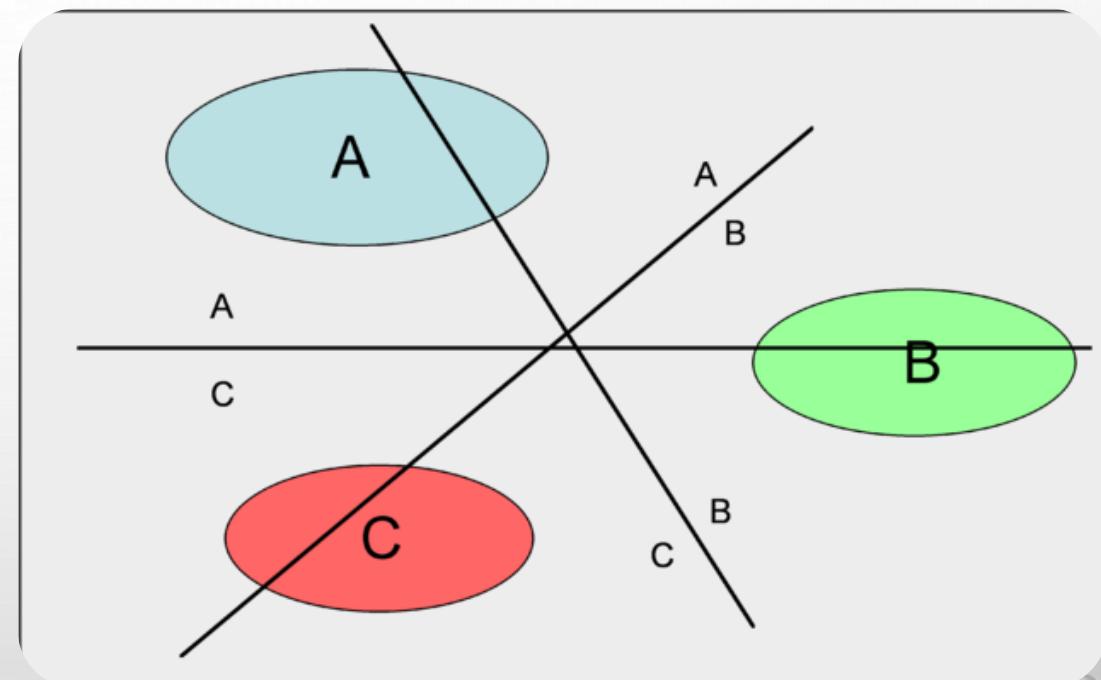
- Discriminar múltiplos objetos em paralelo.
- Ensembles podem ser utilizados para especializar modelos.
- Alguns modelos são naturalmente multiclasse, como redes neurais.



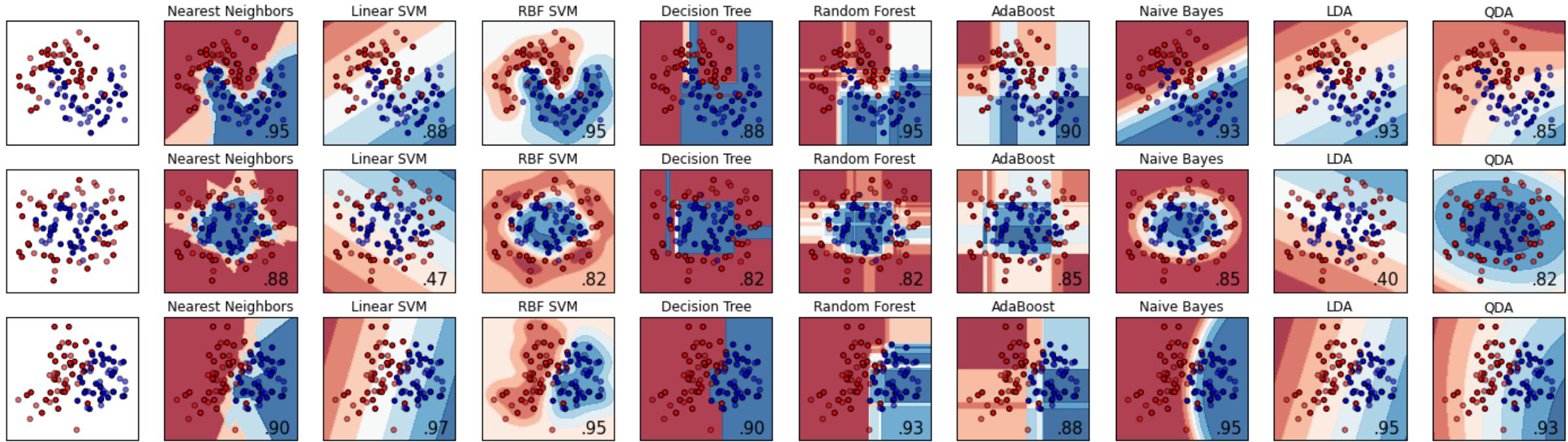
ENSEMBLES BÁSICOS



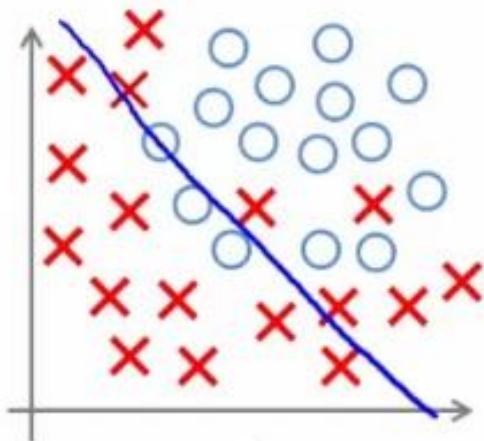
ONE AGAINST ALL



ONE AGAINST ONE

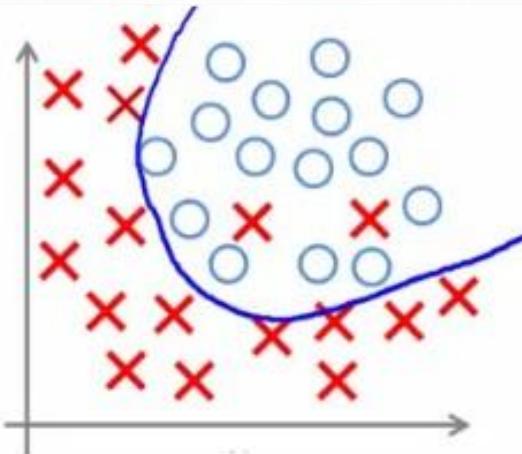


CAPACIDADE E GENERALIZAÇÃO

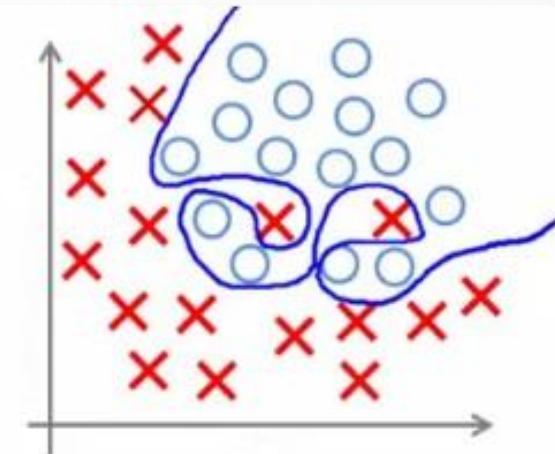


Under-fitting

(too simple to explain the variance)



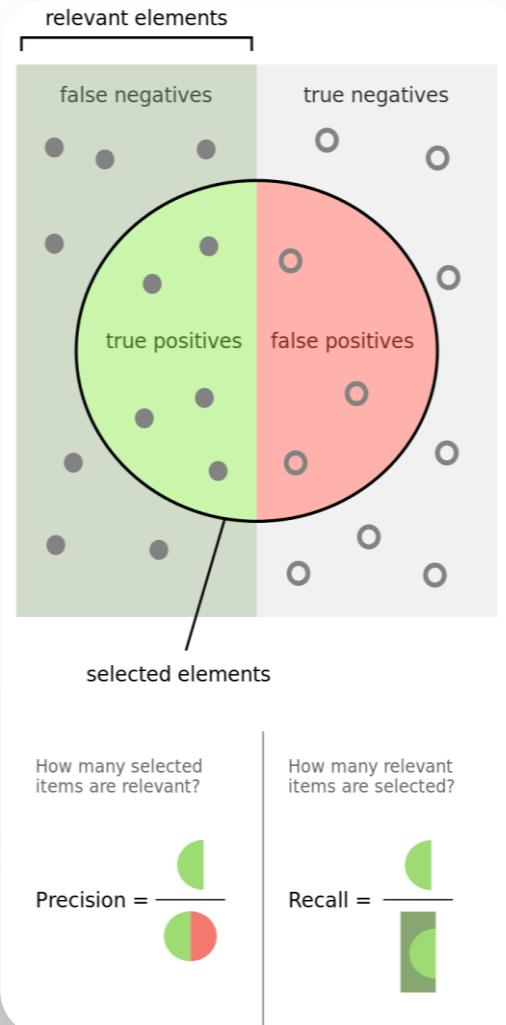
Appropriate-fitting



Over-fitting

(forcefitting – too good to be true)

FIGURAS DE MÉRITO CLASSIFICAÇÃO



Acurácia

- $(TP+TN)/(P+N)$

Taxa de Erro

- 1-Acurácia

Sensibilidade (Recall)

- $TP/(TP+FN)$

Especificidade

- $TN/(TN+FP)$

Precisão

- $TP/(TP+FP)$

Produto Sp

- $\text{SQRT}[\text{SQRT}(R1 * R2) * (R1 + R2)/2]$

GENERALIZAÇÃO: IDENTIFICANDO OS HIPER- PARÂMETROS ÓTIMOS

LEAVE ONE OUT

- Uma única observação é deixada de fora a cada treinamento. N treinamentos são realizados para calcular a estatística de erro.

K FOLDS

- Amostra é dividida em K conjuntos. K treinamentos são realizados, mantendo um conjunto como fora-da-amostra.

BOOTSTRAPPING

- O algoritmo itera, amostrando aleatoriamente M observações, para a quantidade Q desejada de treinamentos.

TREINAMENTO K FOLDS

- TREINAMENTO UTILIZANDO K PARTIÇÕES, COM DADOS DAS CLASSES BALANCEADOS.
- CADA TREINAMENTO É REALIZADO PARA EXPLORAR UMA CONFIGURAÇÃO DE HIPERPARÂMETROS DO MODELO.
- 4 PARTIÇÕES SÃO USADAS PARA TREINAR O MODELO, 1 PARTIÇÃO É UTILIZADA PARA MENSURAR O DESEMPENHO FORA DA AMOSTRA (GENERALIZAÇÃO).
- UMA ESTATÍSTICA DA FIGURA DE MÉRITO É SELECIONADA PARA MEDIR A QUALIDADE DE CADA CONFIGURAÇÃO DE HIPERPARÂMETRO.

Teste

Treino

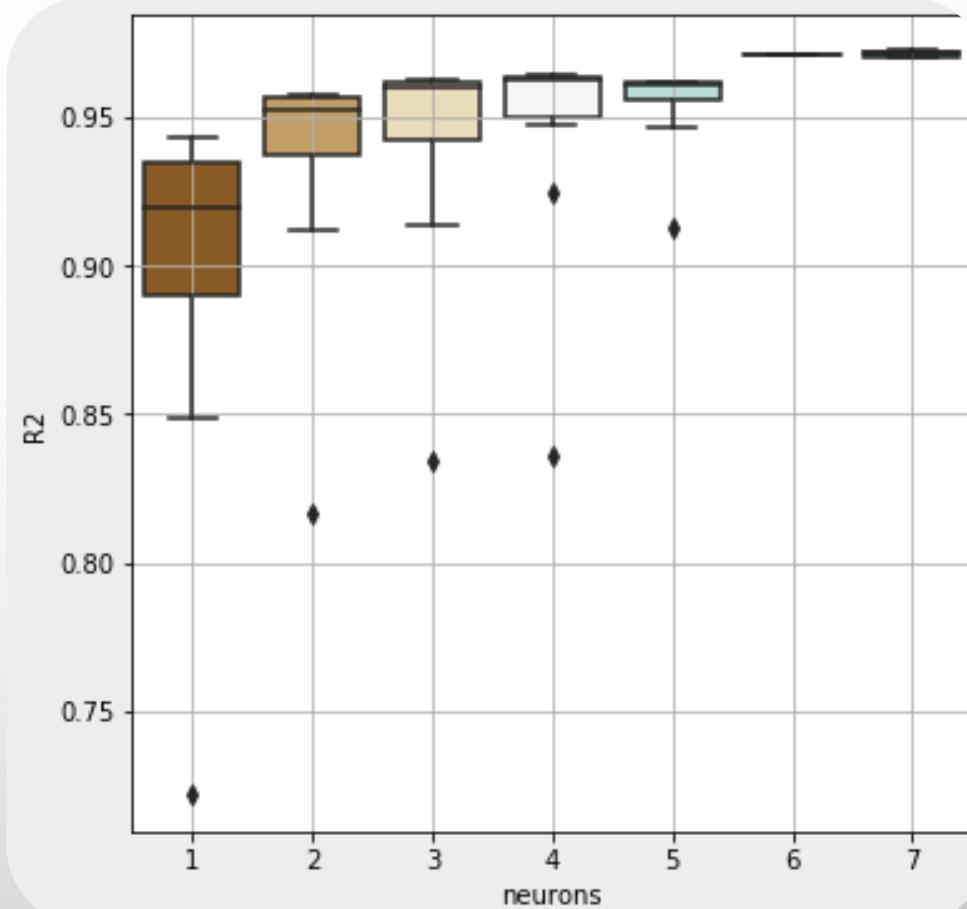
Treino

Treino

Treino

K-FOLDS - EXEMPLO

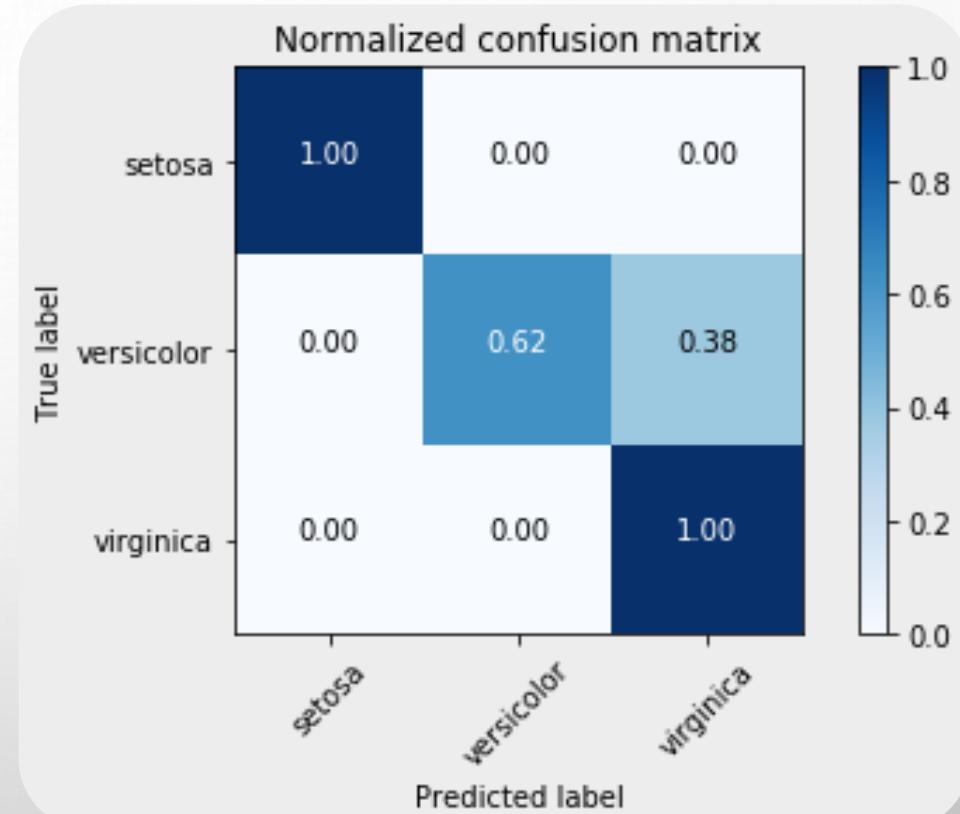
- Iteração dos hiperparâmetros
- Seleção da Figura de Mérito
- Seleção da Estatística de Ganho



EVALUATION

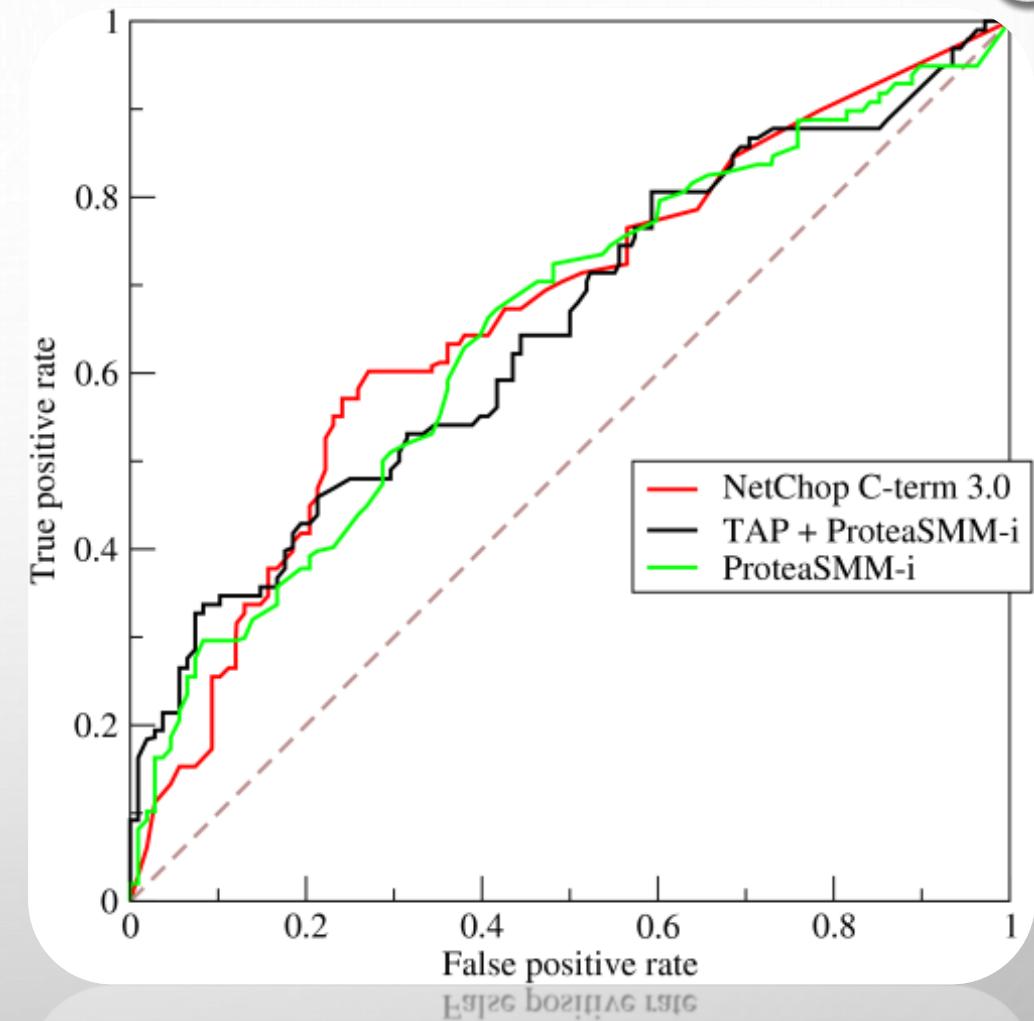
MATRIZ DE CONFUSÃO

Comparação entre o resultado do classificador para as diferentes classes.



VALIDAÇÃO

- **Curva ROC**
 - Calibra a saída do modelo, ajudando a configurar o ponto de operação entre Precisão / Recall / Acurácia.

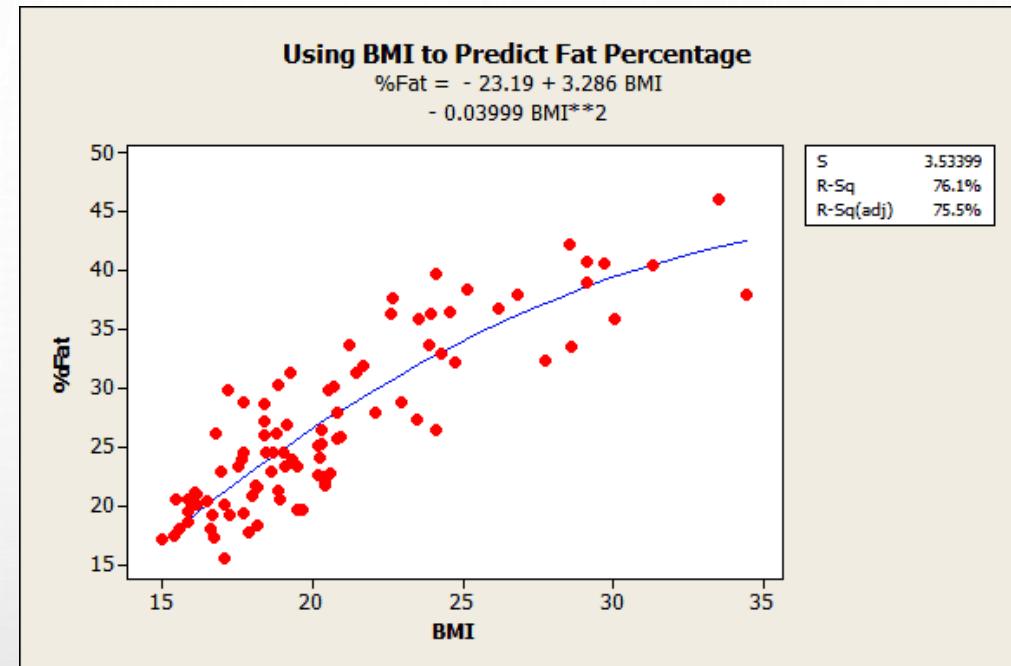


REGRESSÃO

MODELING

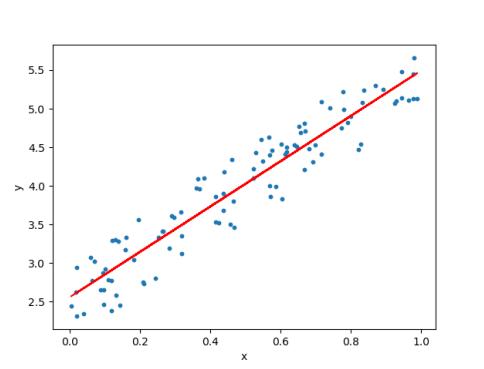
MODELOS DE REGRESSÃO

- 1) Regressão Linear
- 2) Regressão Não-Linear
- 3) Processos Gaussianos
- 4) Máquina de Vetores Suporte
- 5) Redes Neurais

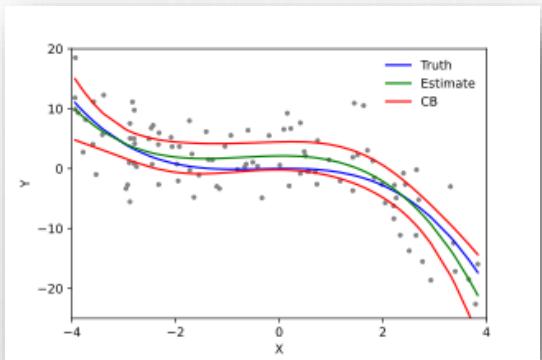


Algoritmos de regressão geralmente são modelados combinando uma **parte determinística e uma parte aleatória**. Os parâmetros correspondente à parte determinística são encontrados utilizando estimadores como máxima verossimilhança ou máximo a posteriori (MAP).

MODELOS DE REGRESSÃO



$$Y = \alpha^T x + \varepsilon$$



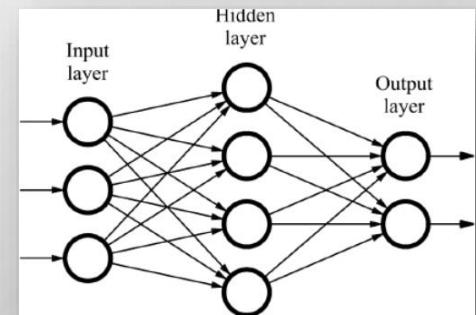
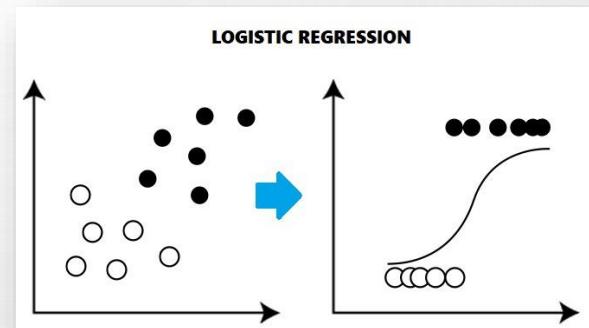
$$Y = X\alpha + \varepsilon$$

$$Y = \frac{1}{1 + e^{\alpha^T x + \varepsilon}}$$

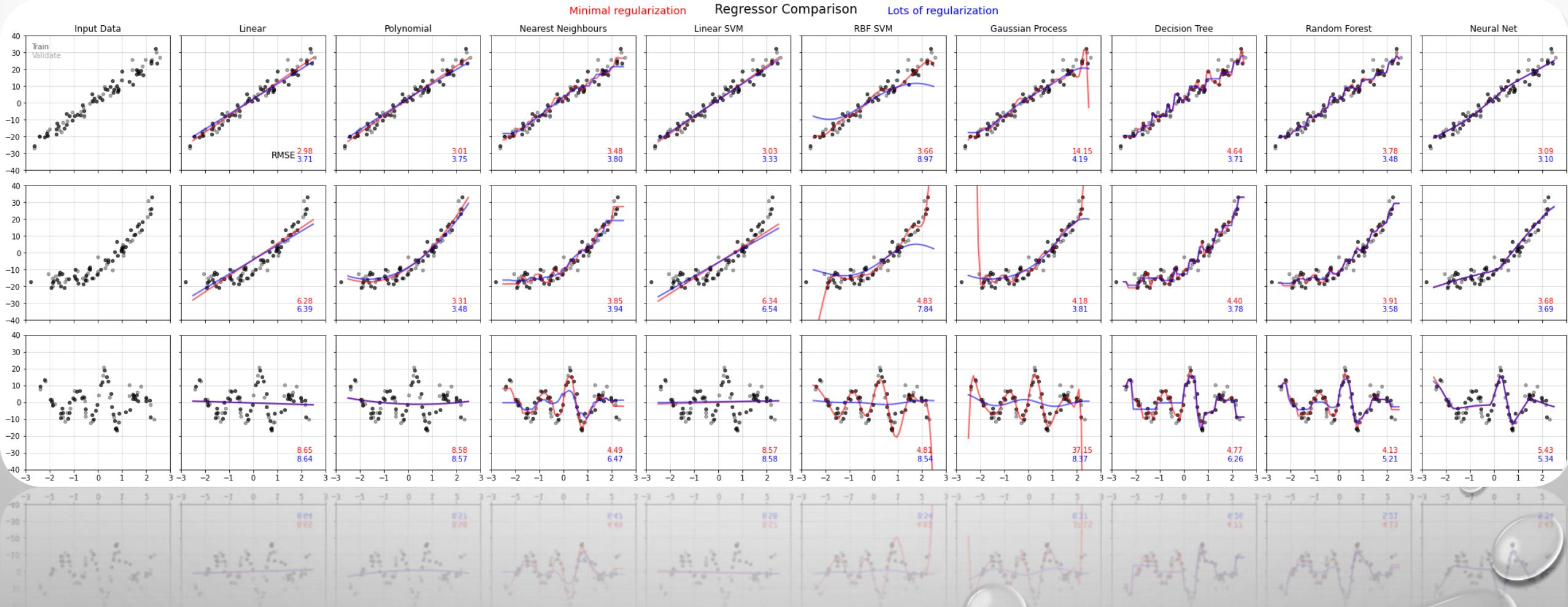
$$Y = F(X) + \varepsilon$$

Parte Determinística

Parte Estocástica



$$Y = \varphi(x) + \varepsilon$$



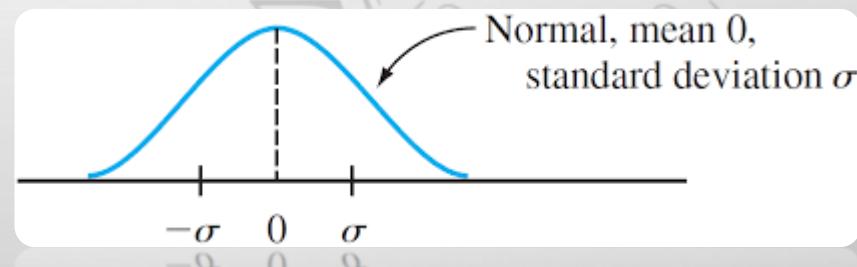
EVALUATION

FIGURAS DE MÉRITO - REGRESSÃO

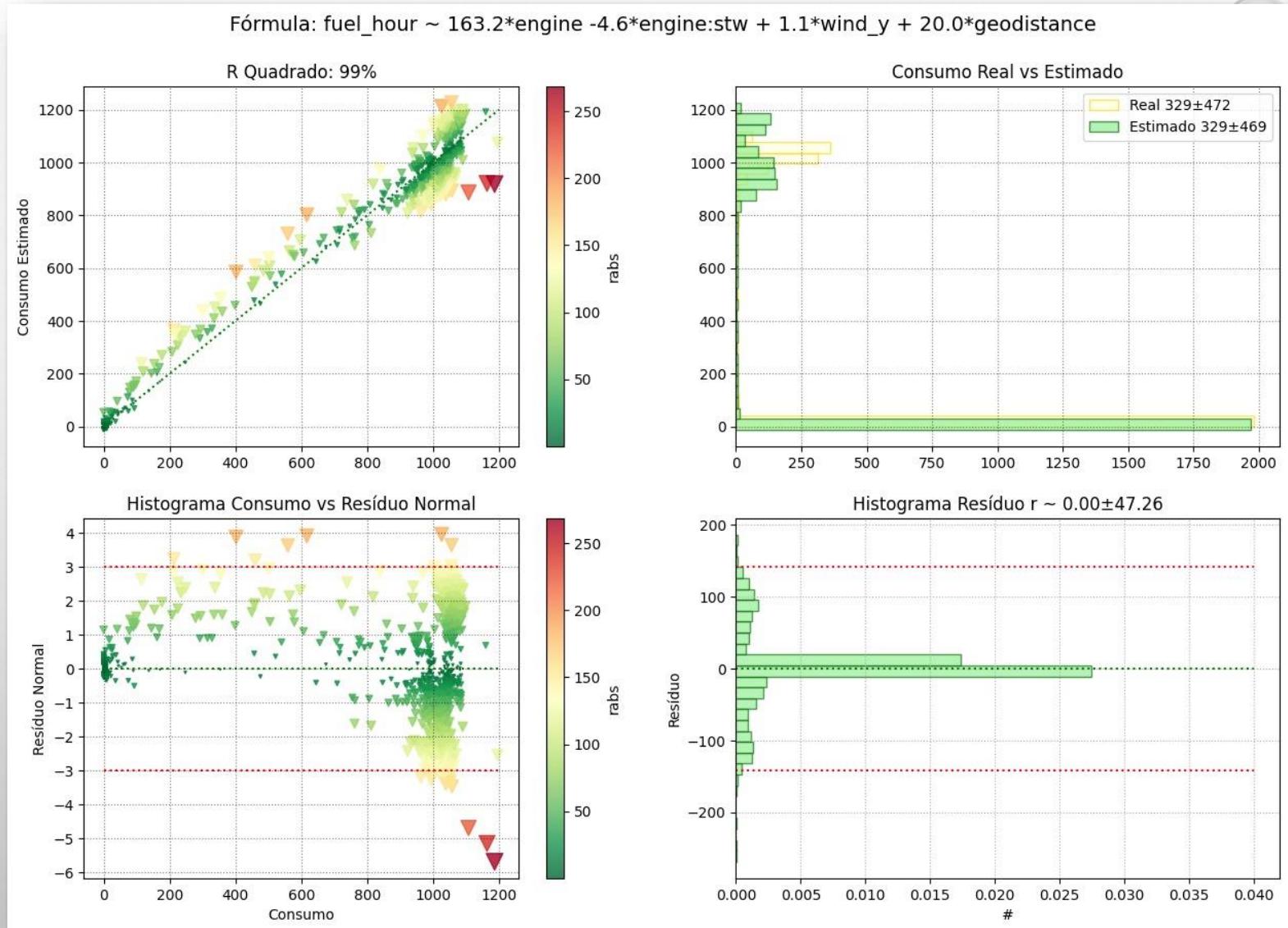
- R QUADRADO

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- RESÍDUO NORMAL DE MÉDIA ZERO E VARIÂNCIA CONSTANTE



ANÁLISE DOS RESULTADOS EXEMPLO



APRENDIZADO NÃO-SUPERVISIONADO

Aprendizado Não-Supervisionado

Não existe um **conhecimento “a priori” dos grupos** contidos nos dados. Algoritmos de agrupamento dependem fortemente de uma definição de “**distância**” ou “**similaridade**” ou “**probabilidade**” entre as observações.

Agrupamento

Um bebê consegue **agrupar objetos por cor, tamanho, formato** e muitos outros atributos que ele pode observar nos objetos.

Diferentes maneiras de organizar os objetos são diferentes **estruturas de agrupamentos** existentes em uma amostra de dados.

APRENDIZADO NÃO-SUPERVISIONADO



De quantas maneiras estes blocos podem ser organizados em grupos?

Um **modelo de agrupamento** é usado para **identificar grupos**, ou estruturas de agrupamentos, nos dados.

Modelagem Probabilística

Em uma cesta de supermercado existe uma variedade grande de itens comprados juntos. A presença de um item específico afeta a **chance de outro produto estar presente na cesta?**

APRENDIZADO NÃO-SUPERVISIONADO

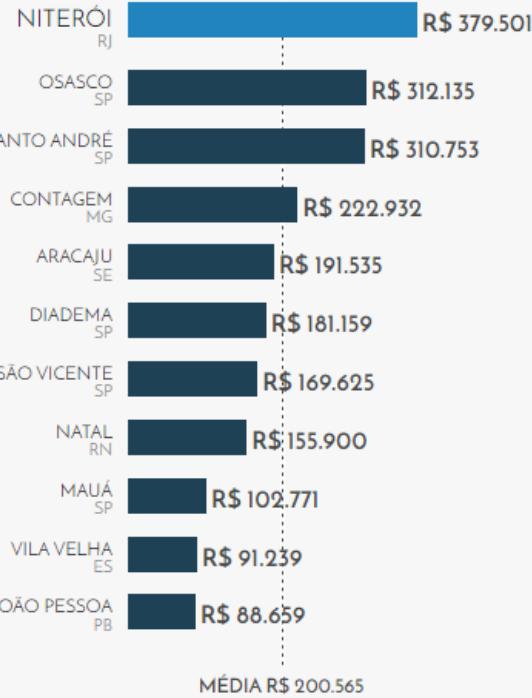


Será que algum desses itens são comprados juntos?

O objetivo de modelos probabilísticos é **identificar padrões** dentro de comportamentos supostamente **aleatórios**.

AGRUPAMENTO

REPRESENTAÇÃO: COMO ENCONTRAR OS 10 MUNICÍPIOS MAIS SIMILARES A NITERÓI?



VARIÁVEIS QUE FORMAM O GRUPO
COMPARAÇÃO

■ Seu município ■ Média do grupo

Domicílios urbanos - (QTD)

169.162	169.822
---------	---------

Características do Entorno

79,34%	73,59%
--------	--------

Domicílios subnormais - (QTD)

24.286	21.725
--------	--------

Renda média domiciliar

R\$ 4.687	R\$ 2.503
-----------	-----------

Saneamento básico - (QTD)

133.750	136.548
---------	---------

*No gráfico ao lado, é possível comparar o município selecionado com os 10 outros municípios brasileiros de perfil mais semelhante para cada item de receita.

Para cada um destes, foi definido o conjunto de variáveis que mais afetam seu resultado – por exemplo, frota de veículos influencia fortemente o valor total de IPVA.

Por meio dos valores dessas variáveis, chega-se aos 10 municípios mais comparáveis com o selecionado.

Veja acima as variáveis que foram utilizadas para o componente de receita definido.

Clique em cada variável acima para entender sua importância.



VARIÁVEIS QUE FORMAM O GRUPO
COMPARAÇÃO

■ Seu município ■ Média do grupo

Domicílios urbanos - (QTD)

169.162	169.822
---------	---------

Características do Entorno

79,34%	73,59%
--------	--------

Domicílios subnormais - (QTD)

24.286	21.725
--------	--------

Renda média domiciliar

R\$ 4.687	R\$ 2.503
-----------	-----------

Saneamento básico - (QTD)

133.750	136.548
---------	---------

*No gráfico ao lado, é possível comparar o município selecionado com os 10 outros municípios brasileiros de perfil mais semelhante para cada item de receita.

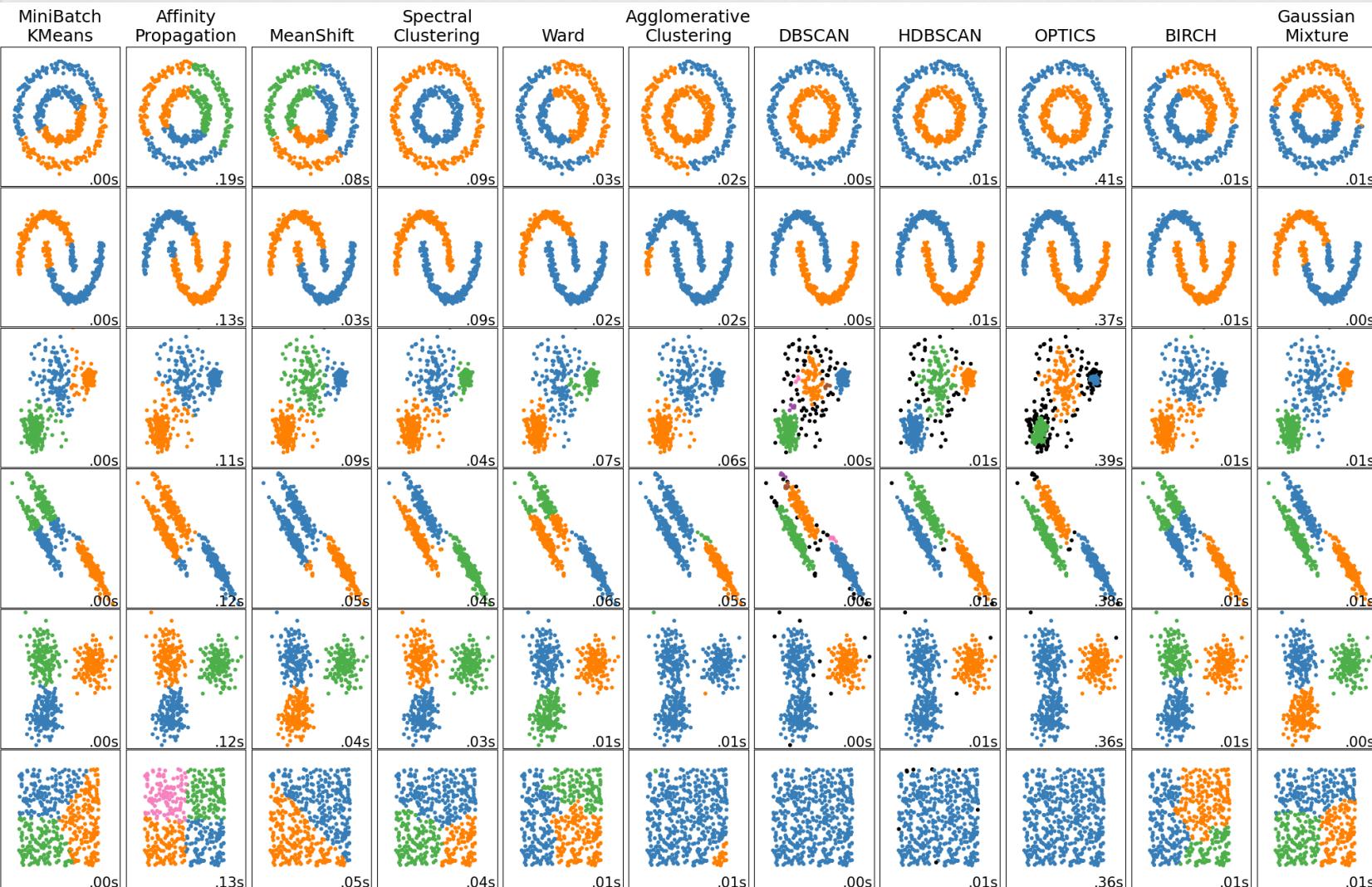
Para cada um destes, foi definido o conjunto de variáveis que mais afetam seu resultado – por exemplo, frota de veículos influencia fortemente o valor total de IPVA.

Por meio dos valores dessas variáveis, chega-se aos 10 municípios mais comparáveis com o selecionado.

Veja acima as variáveis que foram utilizadas para o componente de receita definido.

Clique em cada variável acima para entender sua importância.

MODELAGEM

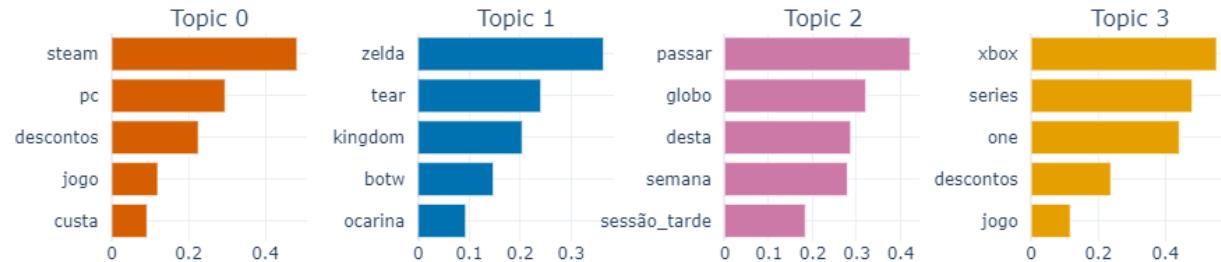


QUANTOS
AGRUPAMENTOS
EXISTEM NESSES
DADOS?

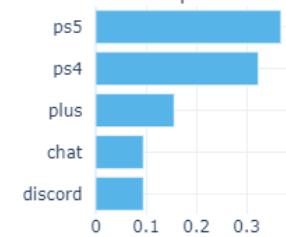
ANÁLISE DE TÓPICOS

MODELAGEM

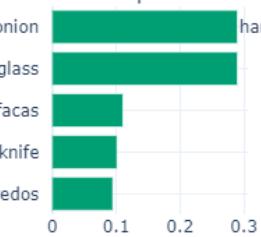
Topic Word Scores



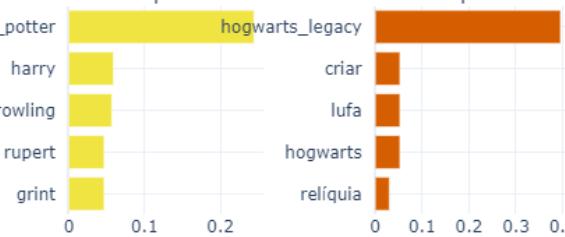
Topic 4



Topic 5



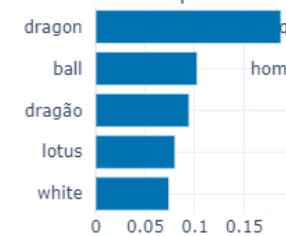
Topic 6



Topic 7



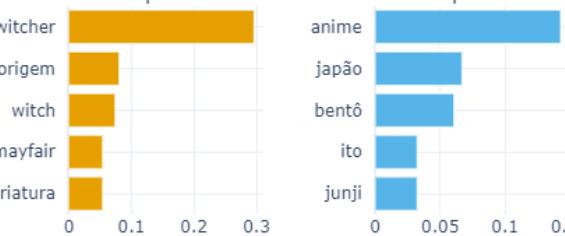
Topic 8



Topic 9



Topic 10



Topic 11



QUAIS TÓPICOS

QUENTES EM REDES

SOCIAIS NERDS?

PERGUNTAS?