
dextra


CRAFTING
SOFTWARE
transforming
BUSINESS

Apache Mahout

API Java de Machine Learning

dextra

O que vai rolar?

O que é Mahout?

Como um sistema pode se beneficiar de ML?

Recomendação

Classificação / Predição

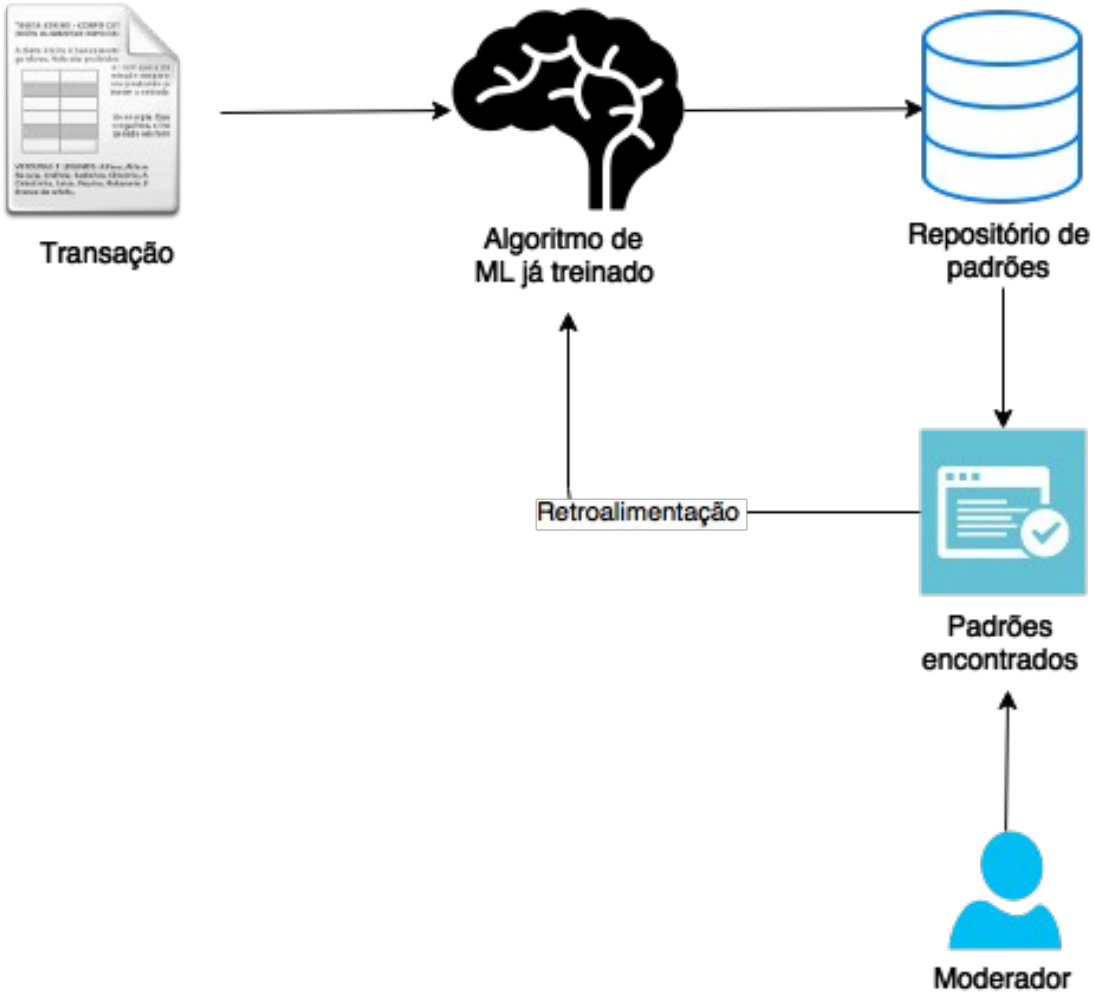
Segmentação

:)

Como funciona a Classificação?

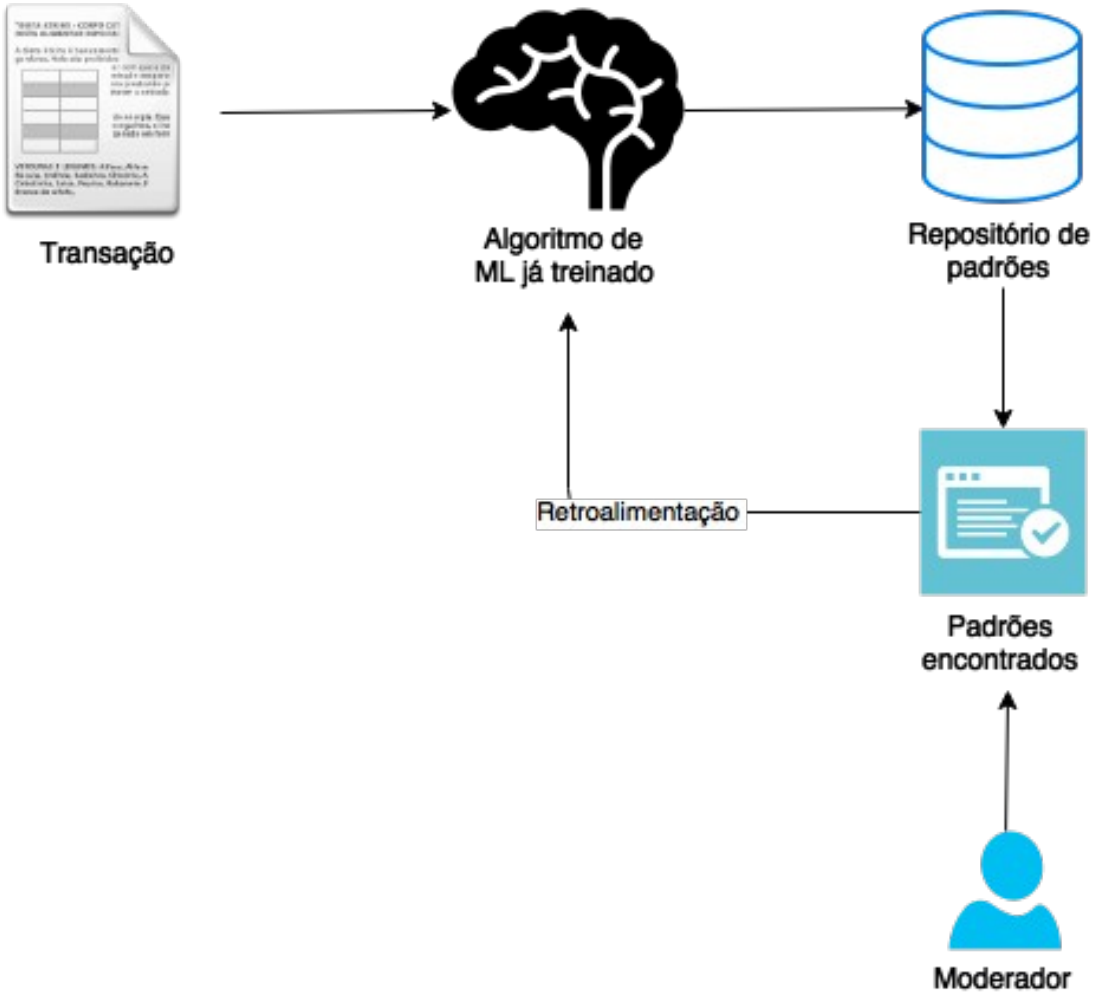
Como funciona a classificação?

Ex: Classificar uma transação clinica como “Fraude” ou “Não-Fraude”.



Como funciona a classificação?

Ex: Classificar uma transação clinica como “Fraude” ou “Não-Fraude”.



Algoritmo **JÁ TREINADO**!

Como treinar um algoritmo?

Precisa de exemplos de dados!

CLASSE	TP_MOVIM.	CD_ESPEC.	CD_PREST.	RESULTADO
TRIVIAL	10	320	10	**0**
TRIVIAL	16	311	12	**0**
ANOMALIA	14	598	35	**1**
ANOMALIA	20	613	29	**1**

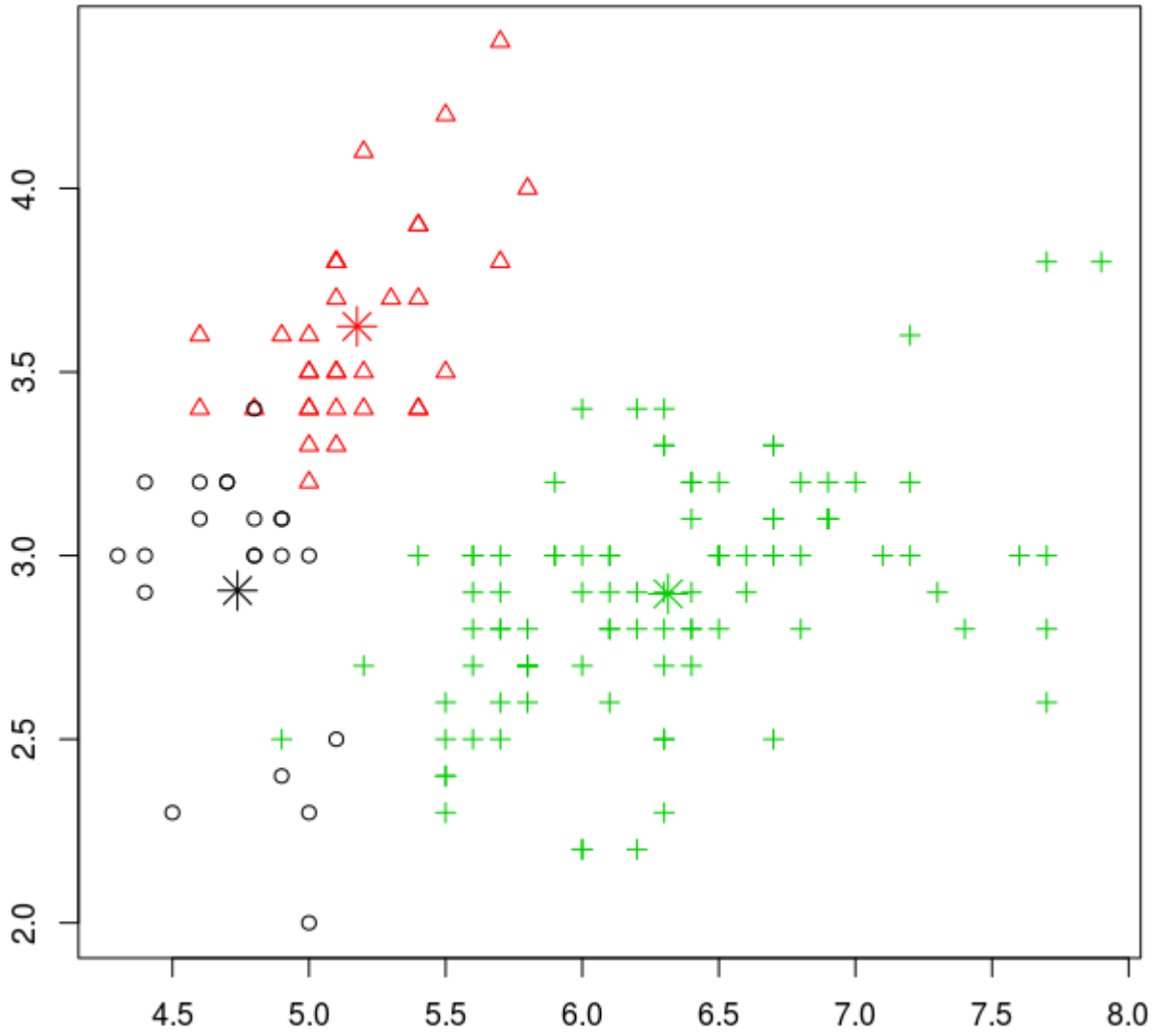
Código |o|

Blz! Como separar anomalias de trivialidades?


AGORA FUDEU

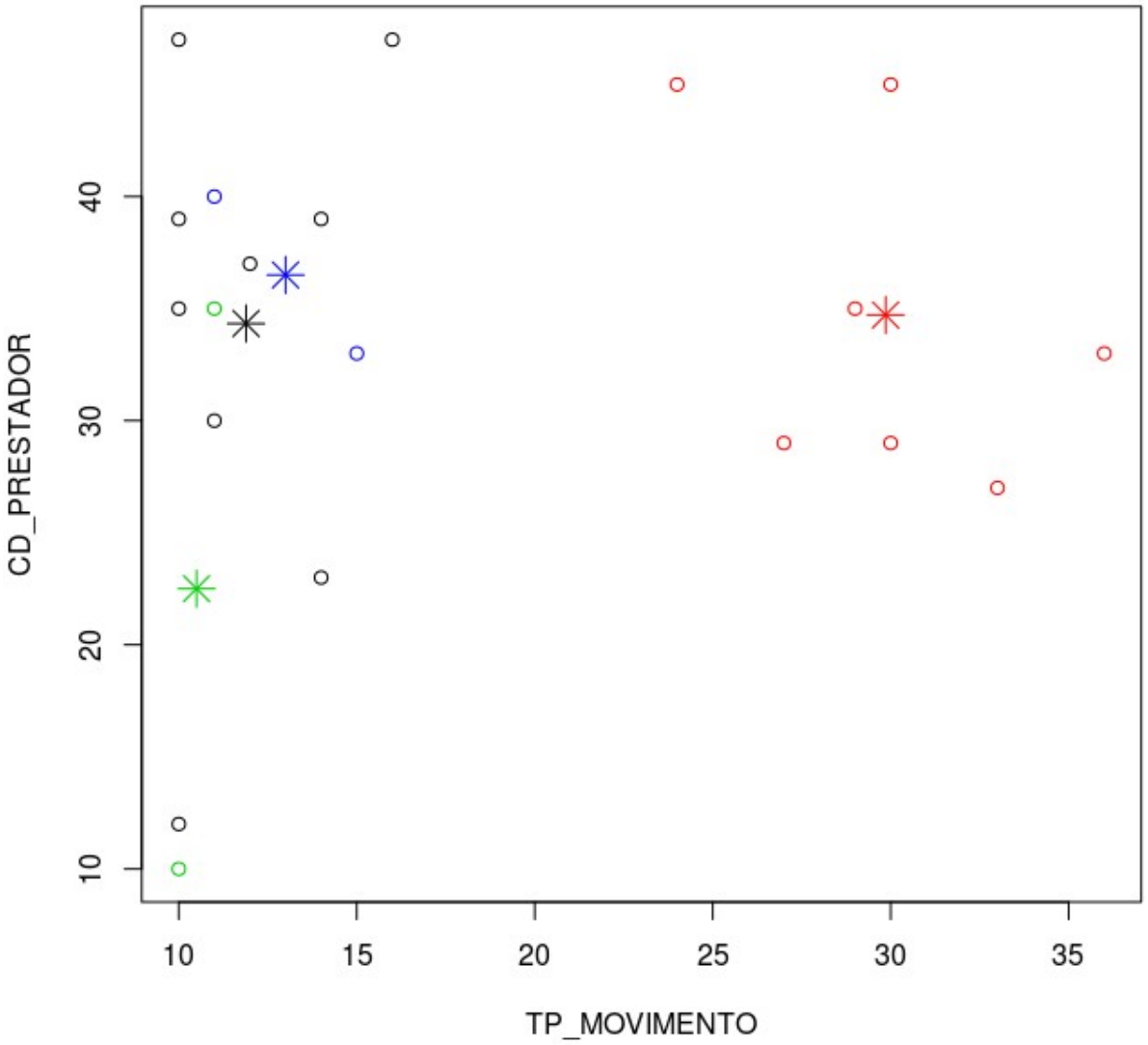
Dividir para conquistar... Vamos Segmentar!

Segmentação:



Código |o|

Segmentação:



Problemas na segmentação:

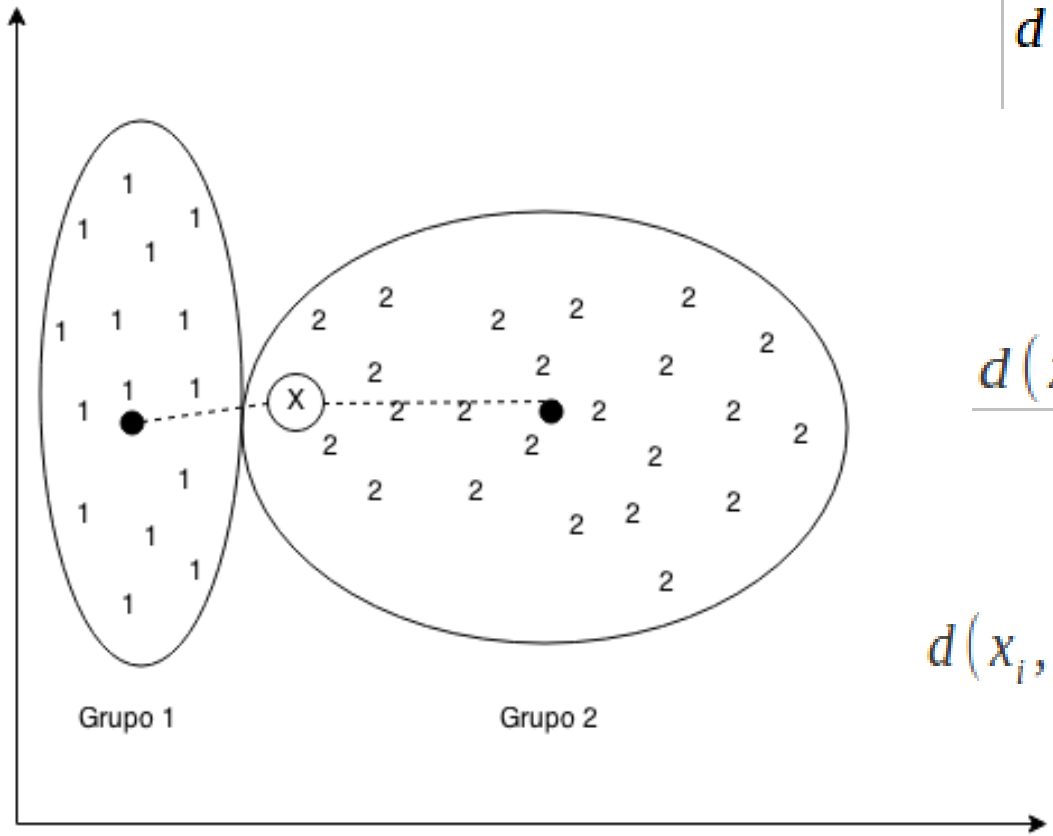
- Posicionamento inicial dos centroides;
- Métrica de distância;
- Número de clusters é parâmetro de entrada.

Posicionamento dos centroides:

- Aleatório;
- Controlado:
  - Por triangulação de dispersão;
  - Densidades de massa por estimativa de gradiente.

Métrica de distância:

- Euclidiana vs Mahalanobis?



$$d(x_i, x_j) = [\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2]^{\frac{1}{2}}$$

$$d(x_i, y_i) = \frac{[\sum_{k=1}^{d} (x_{i,k} - y_{i,k})^2]^{\frac{1}{2}}}{\sigma}$$

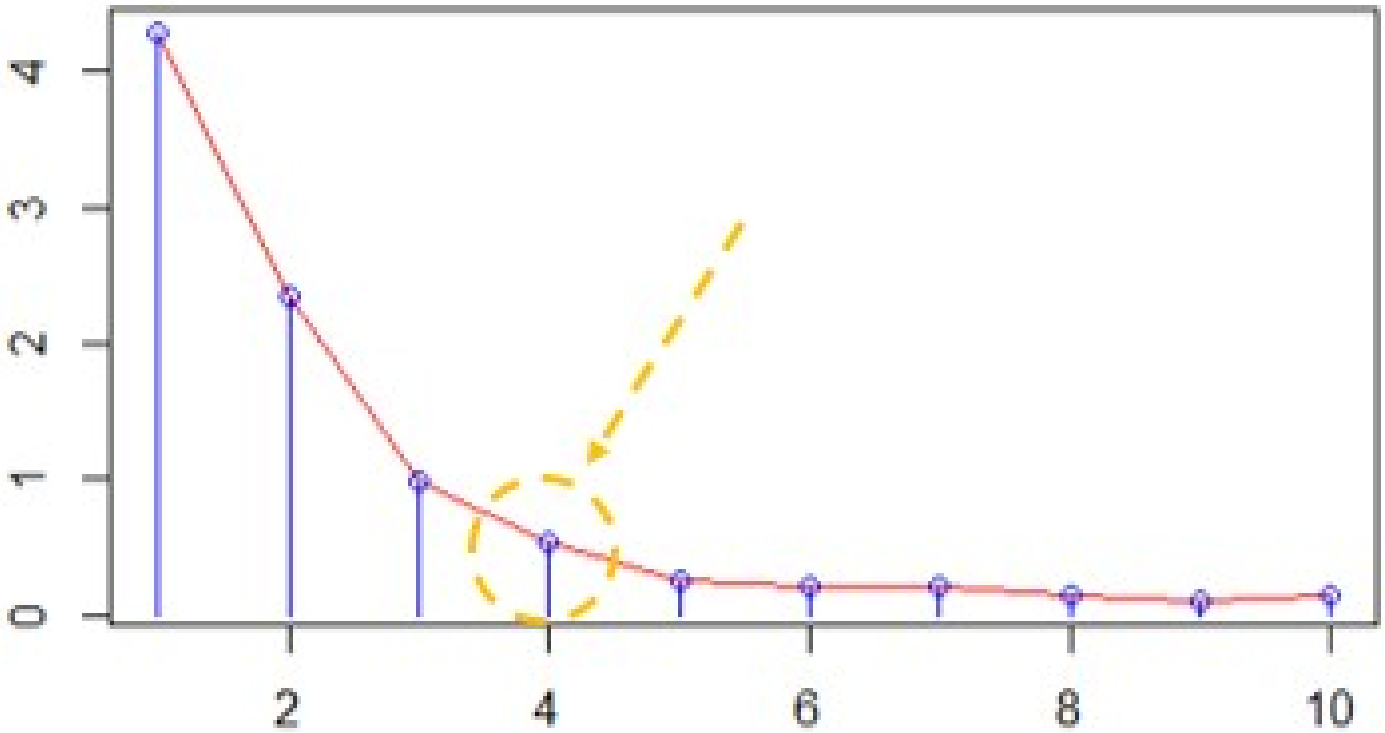
$$d(x_i, y_i) = [\sum_{k=1}^{d} (\bar{x} - \bar{y})^T A^{-1} (\bar{x} - \bar{y})]^{\frac{1}{2}}$$

dextra

Número de clusters:

- Elbow Method.

$$D_k = \sum_{x_i \in C_k} \sum_{x_j \in C_k} ||x_i - x_j||^2 = 2n_k \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$



Case

Case: Segmentação dos dados.

- Não tinhamos uma separação clara de exemplos de anomalias e trivialidades.
- Segmentamos (agrupamos) os dados para melhorar o entendimento. Resultado: Muito trabalho para analisar.
- Concluímos que os procedimentos de glosa deveriam ser marcados como anomalias.
- Descoberta de alguns comportamentos “curiosos” :)

Case: Classificação / Predição.

- Alguns alarmes de falso positivo no início, mas o mecanismo de retroalimentação melhora gradativamente a acurácia do classificador.
- O treinamento do classificador não é on-line. Um *job* agendado realiza o treinamento periodicamente.
- O “modelo de ação” sobre o negócio mudou. É diferente quando encontramos um problema antes do cliente Preventivo vs Reativo :)

Case: Mahout.

- Escalabilidade (map reduce).
- API madura e com suporte efetivo.
- Flexível e suporta mudanças / personalizações.
- A única forma persistir o “conhecimento” adquirido com o treinamento é serializando o objeto treinado no banco :(

Cuidados

Cuidados:

Classification	Single Machine	MapReduce
Logistic Regression	x	
Naive Bayes		x
Random Forest		x
Multilayer Perceptron	x	
Clustering	**Single Machine**	**MapReduce**
k-Means Clustering	x	x
Fuzzy k-Means	x	x
Spectral Clustering		x

Cuidados:

- Treinamento do classificador:
  - Separar seu conjunto de dados entre treino e teste (*Cross Validation*). Essa é uma forma de avaliar a capacidade de generalização do classificador.
  - Eliminar redundâncias do conjunto de treino;

Obrigado!

everton.gago@dextra-sw.com