Aprendizagem de Máquina aplicada à saúde com R e mlr³

Seleção de características à interpretação de modelos.

Prof. Patrick Terrematte - UFERSA

Doutorando em Bioinformática - BioME/UFRN

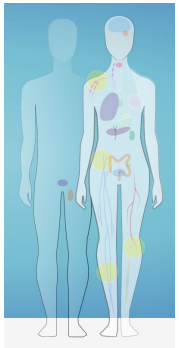
patrick.terrematte@ufersa.edu.br

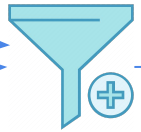
Tópicos

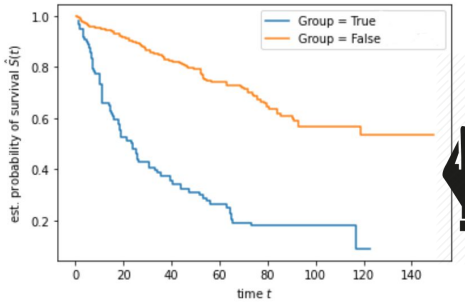
1. Introdução
2. Construíndo modelos
    a. Treinamento, Predição e Validação
    b. Benchmarking
3. Otimização de modelos
    a. Ajuste de hiperparâmetros
    b. Seleção de características
4. Pipelines
5. Hands-on tutorial

1. Introdução

Seleção de assinatura gênica para sobrevida







Integração de dados

- Kidney clear cell - KIRC

Seleção de dados clínicos, mutações e expressão gênica

- Conjunto mínimo de genes relevantes

Aprendizagem de máquina e tomada de decisão

- Análise de sobrevida

- Prof. Dr. Adrião Duarte, BioME / Ppgee
- Profa. Dra. Beatriz Stransky, BioME / Eng. Biomédica
- Patrick Terrematte, Doutorando BioME

Assinaturas de genes

> Clin Cancer Res. 2005 Aug 15;11(16):5730-9. doi: 10.1158/1078-0432.CCR-04-2225.

Gene signatures of progression and metastasis in renal cell cancer

Jon Jones [1], Hasan Otu, Dimitrios Spentzos, Shakirahmed Kolia, Mehmet Inan, Wolf D Beecken, Christian Fellbaum, Xuesong Gu, Marie Joseph, Allan J Pantuck, Dietger Jonas, Towia A Libermann

Affiliations + expand

PMID: 16115910 DOI: 10.1158/1078-0432.CCR-04-2225

> Medicine (Baltimore). 2018 Nov;97(44):e12679. doi: 10.1097/MD.0000000000012679.

Comprehensive assessment gene signatures for clear cell renal cell carcinoma prognosis

Peng Chang [1 2 3], Zhitong Bing [3 4], Jinhui Tian [3 4], Jingyun Zhang [3 4], Xiuxia Li [3 4 5], Long Ge [3 4], Juan Ling [3 4], Kehu Yang [1 4], Yumin Li [1 2]

Affiliations + expand

PMID: 30383629 PMCID: PMC6221654 DOI: 10.1097/MD.0000000000012679

> J Cell Physiol. 2019 Jul;234(7):10324-10335. doi: 10.1002/jcp.27700. Epub 2018 Nov 11.

Prognostic value of a gene signature in clear cell renal cell carcinoma

Liang Chen [1], Yongwen Luo [1], Gang Wang [2 3], Kaiyu Qian [2 3], Guofeng Qian [4], Chin-Lee Wu [5], Han C Dan [6], Xinghuan Wang [1], Yu Xiao [1 2 3]

> Oncotarget. 2016 Dec 13;7(50):82712-82726. doi: 10.18632/oncotarget.12631.

A four-gene signature predicts survival in clear-cell renal-cell carcinoma

Jun Dai [1], Yuchao Lu [2], Jinyu Wang [3], Lili Yang [4], Yingyan Han [1], Ying Wang [4], Dan Yan [5], Qiurong Ruan [4], Shaogang Wang [2]

> Hereditas. 2020 Sep 3;157(1):38. doi: 10.1186/s41065-020-00152-y.

A seven-gene signature model predicts overall survival in kidney renal clear cell carcinoma

Ling Chen [1], Zijin Xiang [2], Xueru Chen [2], Xiuting Zhu [2], Xiangdong Peng [3]

Affiliations + expand

PMID: 32883362 PMCID: PMC7470605 DOI: 10.1186/s41065-020-00152-y

> PeerJ. 2020 Oct 29;8:e10183. doi: 10.7717/peerj.10183. eCollection 2020.

A 14 immune-related gene signature predicts clinical outcomes of kidney renal clear cell carcinoma

Yong Zou [1], Chuan Hu [1]

Affiliations + expand

PMID: 33194402 PMCID: PMC7603789 DOI: 10.7717/peerj.10183

> Sci Rep. 2020 Feb 6;10(1):2026. doi: 10.1038/s41598-020-58804-y.

Identification of gene signature for treatment response to guide precision oncology in clear-cell renal cell carcinoma

Ninadh M D'Costa [1 2], Davide Cina [1], Raunak Shrestha [1 2], Robert H Bell [1 2], Yen-Yi Lin [1 2], Hossein Asghari [2 3], Cesar U Moniaras-Avila [1 2], Christian Kollmannsberger [2 4], Faraz Hach [1 2]

> Int J Mol Sci. 2019 Nov 14;20(22):5720. doi: 10.3390/ijms20225720.

A Gene Signature of Survival Prediction for Kidney Renal Cell Carcinoma by Multi-Omic Data Analysis

Fuyan Hu [1], Wenying Zeng [2], Xiaoping Liu [3]

> Anim Cells Syst (Seoul). 2020 May 12;24(3):160-170. doi: 10.1080/19768354.2020.1760932.

Construction and validation of a seven-gene signature for predicting overall survival in patients with kidney renal clear cell carcinoma via an integrated bioinformatics analysis

Huiming Jiang [1], Haibin Chen [2], Nanhui Chen [1]

Affiliations + expand

PMID: 33209196 PMCID: PMC7651852 DOI: 10.1080/19768354.2020.1760932

> Front Oncol. 2019 Mar 19;9:152. doi: 10.3389/fonc.2019.00152. eCollection 2019.

Construction and Validation of a 9-Gene Signature for Predicting Prognosis in Stage III Clear Cell Renal Cell Carcinoma

Junlong Wu [1 2], Shengming Jin [1 2], Weijie Gu [1 2], Fangning Wan [1 2], Hailiang Zhang [1 2], Guohai Shi [1 2], Yuanyuan Qu [1 2], Dingwei Ye [1 2]

Affiliations + expand

PMID: 30941304 PMCID: PMC6433707 DOI: 10.3389/fonc.2019.00152

Comment > Nat Rev Nephrol. 2019 Sep;15(9):528. doi: 10.1038/s41581-019-0179-

Gene signatures reveal kidney immune cells

Monica Wang [1]

Affiliations + expand

PMID: 31285592 DOI: 10.1038/s41581-019-0179-7

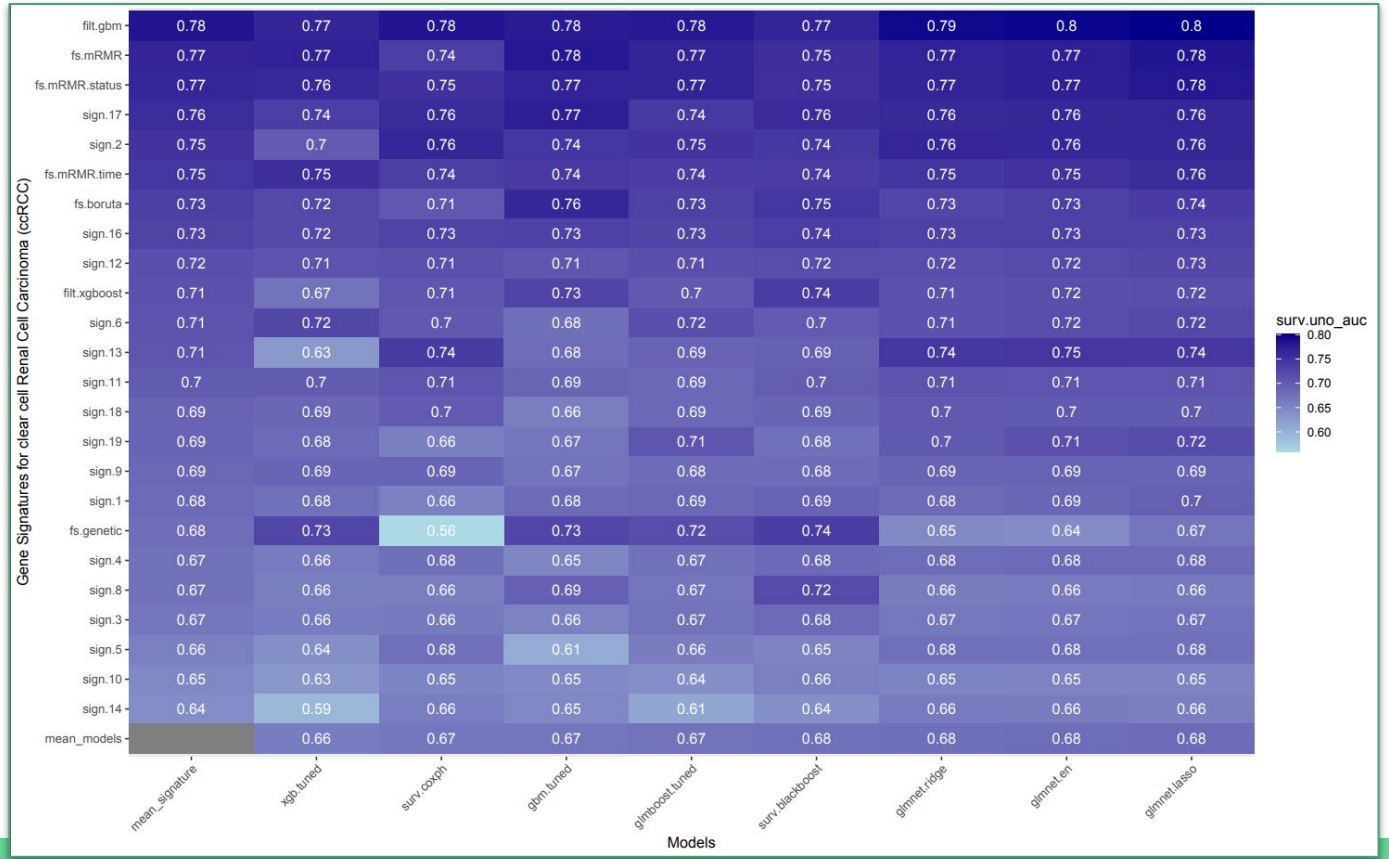
> Med Sci Monit. 2019 Jun 13;25:4401-4413. doi: 10.12659/MSM.917399.

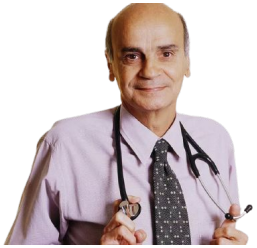
Identification of a 5-Gene Signature Predicting Progression and Prognosis of Clear Cell Renal Cell Carcinoma

Qiufeng Pan [1], Longwang Wang [2], Hao Zhang [1], Chaoqi Liang [1], Bing Li [1]

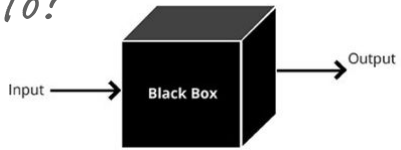
Como escolher uma assinatura adequada?

Benchmark das assinaturas

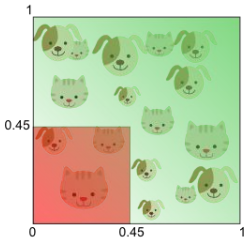


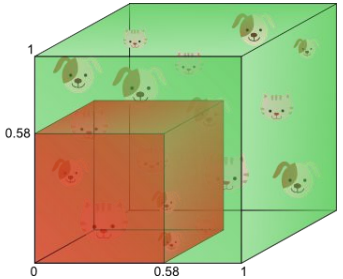


- Vou entender o modelo e decidir o tratamento?
  - Interpretabilidade do Modelo

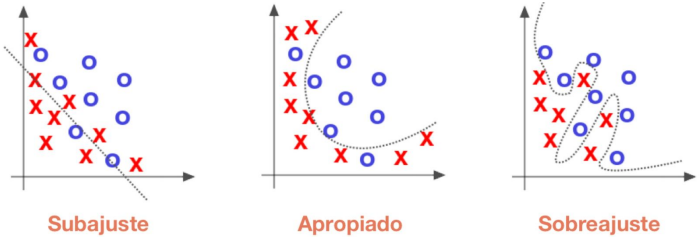


- Vai consumir muita energia treinar o modelo?
  - Maldição da dimensionalidade





- Vou poder usar o modelo e me tratar?
  - Sobreajuste e generalização



CLEANING THE DATA

FITTING THE MODEL

SELECTING FEATURES


EVER
SUEZ CANAL AUTHORIT
INTERPRETING
THE MODEL
EVER
SELECTING
FEATURES
EVER
CLEANING
THE DATA

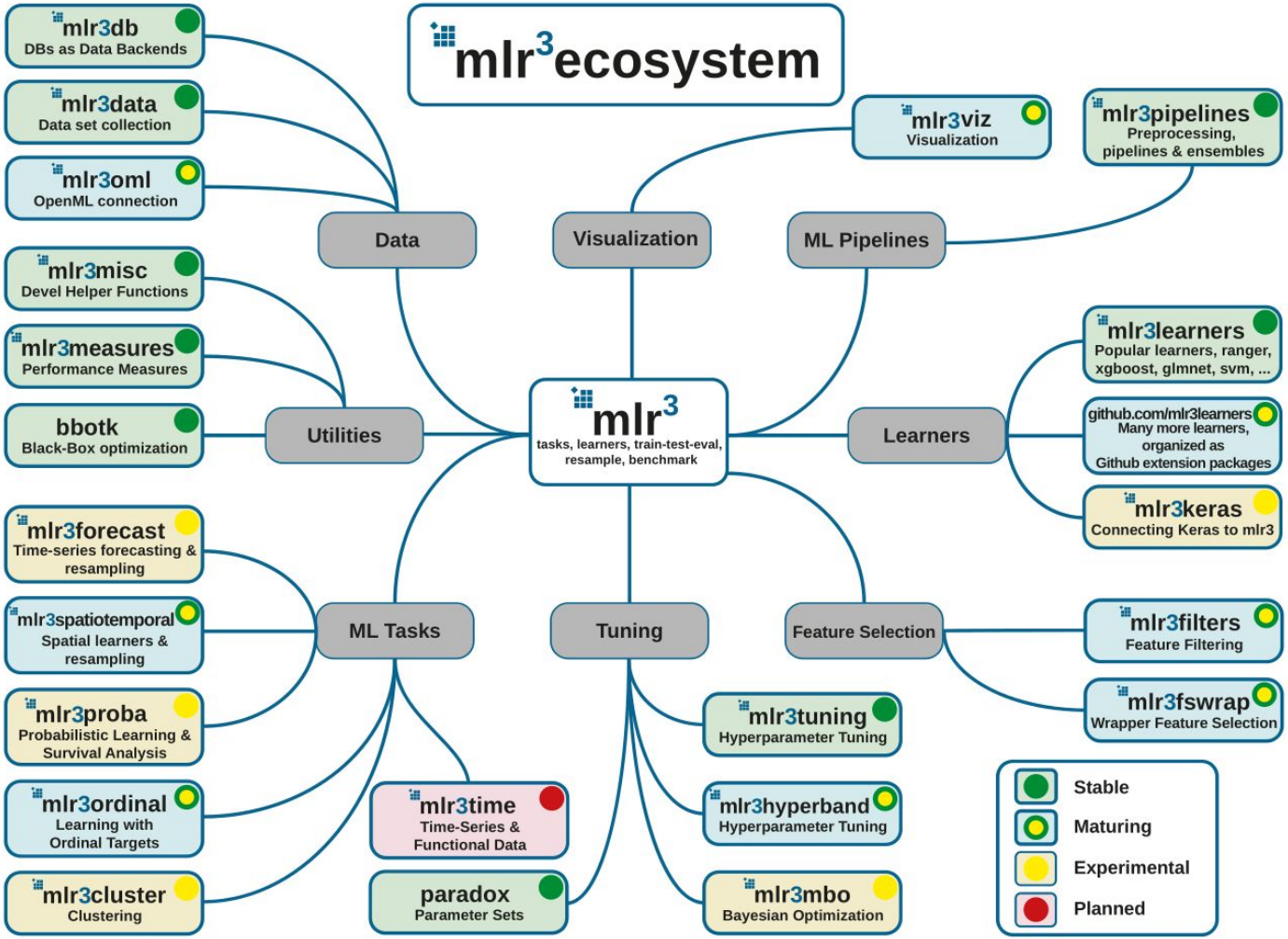
2. Construíndo modelos

Aprendizagem de Máquina em R

- Problema:
  - Na linguagem R há diferentes modelos, mas não há interface unificada

```r
# Definindo um modelo com formula: target ~ features
svm_model = e1071::svm(id=metastasis ~ . , data = kirc_data)
```

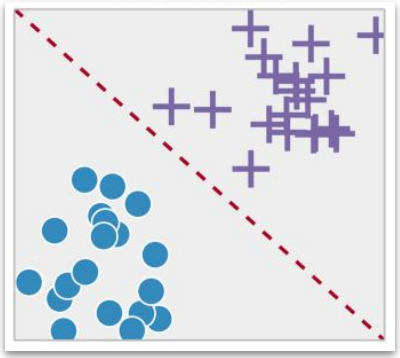
```r
# Passamos features como uma matrix, e o target como vector
xgb_model = xgboost::xgboost(data = as.matrix(kirc_data[,1:9]), label = kirc_data$metastasis)
```

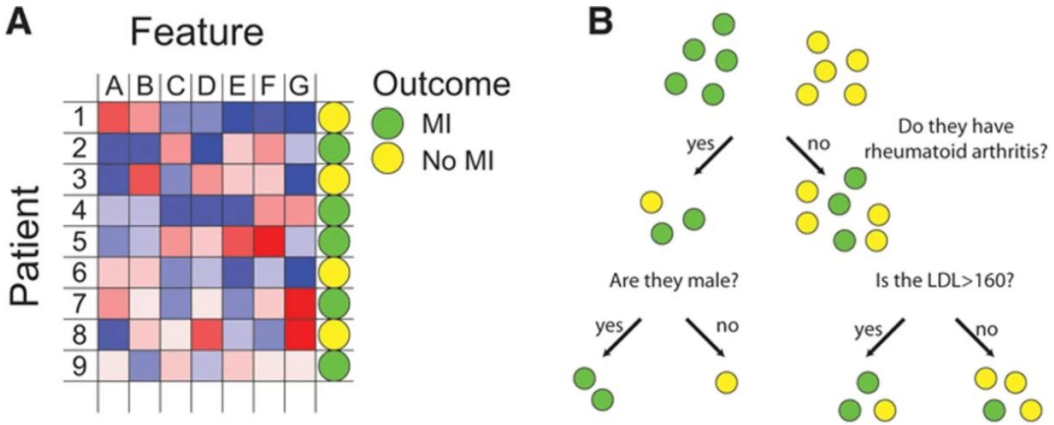


https://mlr3book.mlr-org.com

Classificação

```
tsk_cla = TaskClassif$new(id="kirc_cla",
                          backend = kirc_data,
                          target = "metastasis",
                          positive = "M1")
```

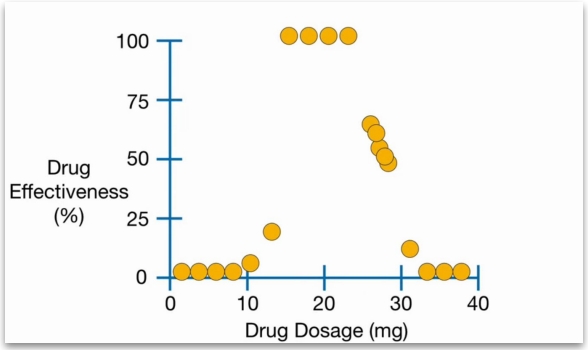


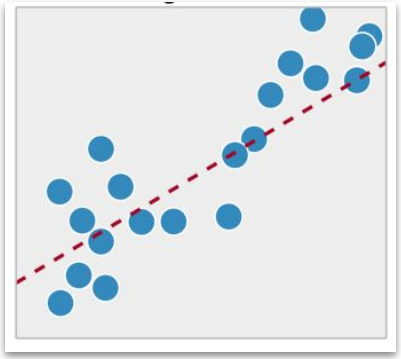


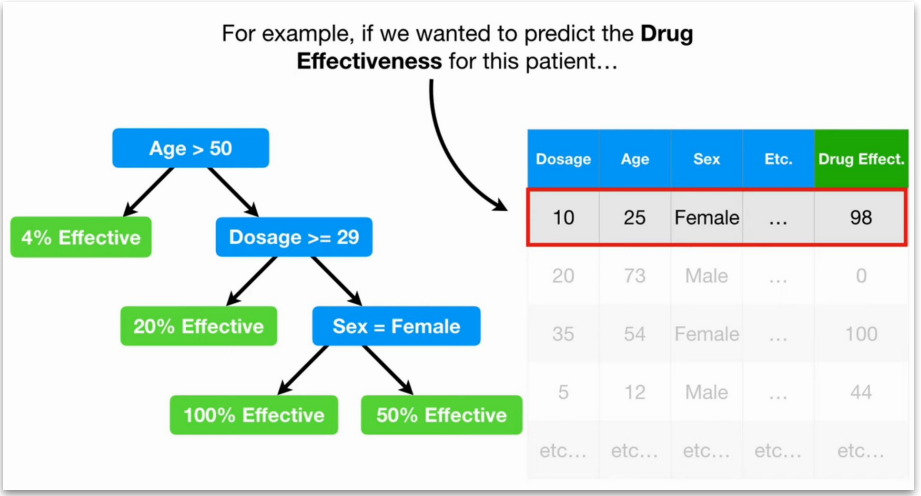
https://emerj.com/ai-sector-overviews/machine-learning-in-pharma-medicine/

Regressão

```
tsk_rgr = TaskRegr$new(id="kirc_rgr",
                       backend = kirc_data,
                       target = "event_time")
```



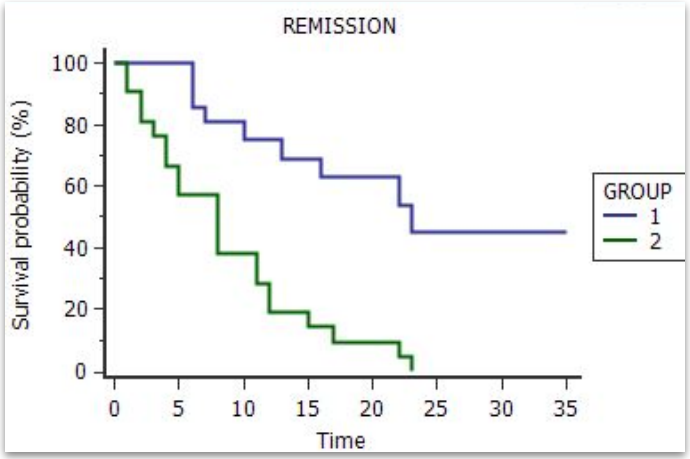


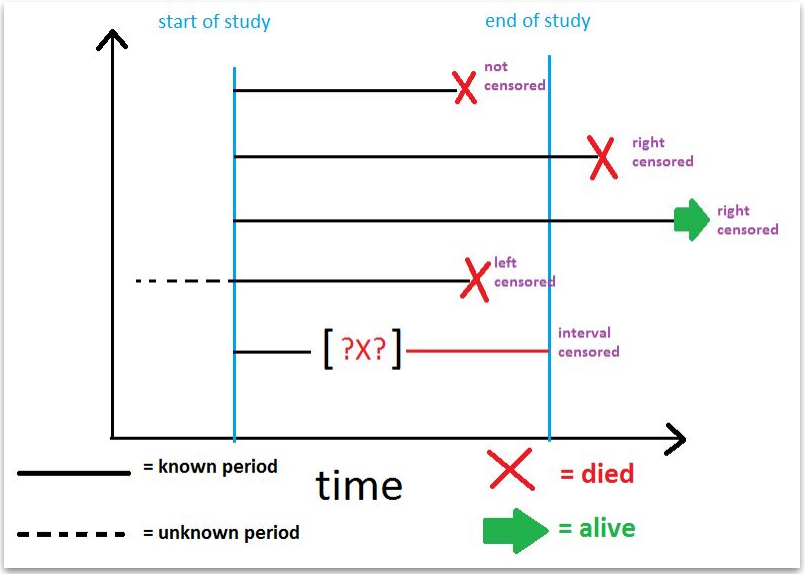


StatQuest with Josh Starmer https://www.youtube.com/watch?v=g9c66TUylZ4

Análise de sobrevida

```
tsk_srv = TaskSurv$new(id="kirc_srv",
                       backend = kirc_data,
                       time = "event_time",
                       event = "obs_death",
                       type = "right")
```

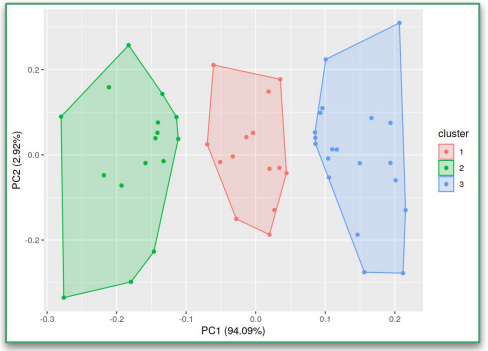




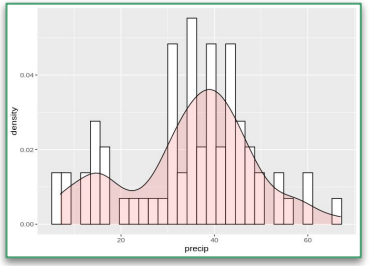
https://www2.karlin.mff.cuni.cz/~pesta/NMFM404/survival.html

Outros tipos de problemas

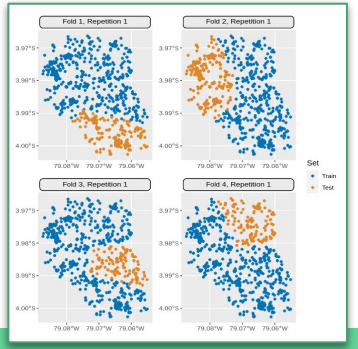
- Agrupamento: **tarefa não supervisionada** para identificar grupos em um espaço de características
  → `mlr3cluster::TaskClust`



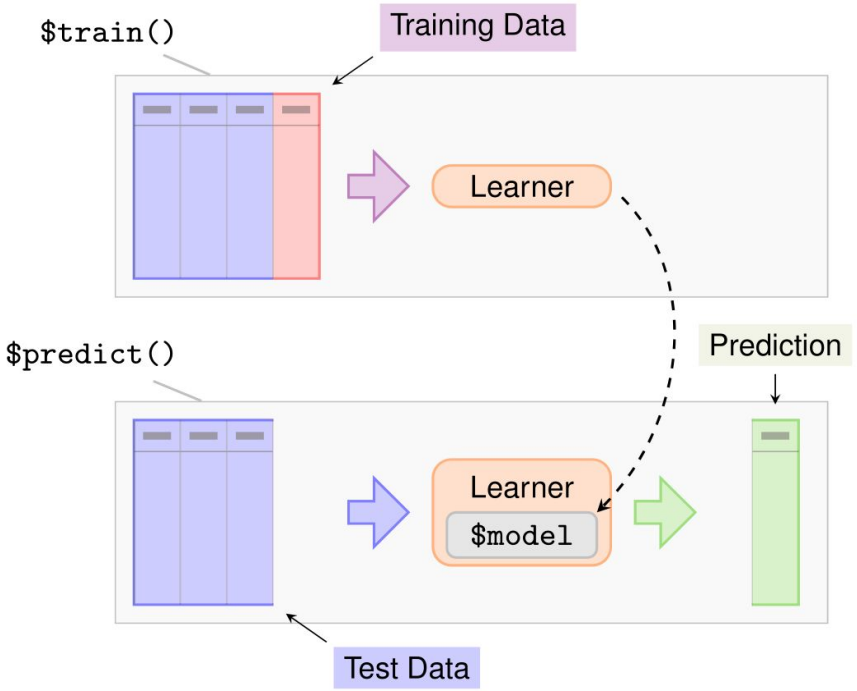
- Densidade: **tarefa não supervisionada** para estimar densidade de distribuições.
  → `mlr3proba::TaskDens`



- Espaço-temporal: **tarefa** sobre dados de coordenadas.
  → `mlr3spatiotempcv::TaskRegrST`
  → `mlr3spatiotempcv::TaskClassifST`

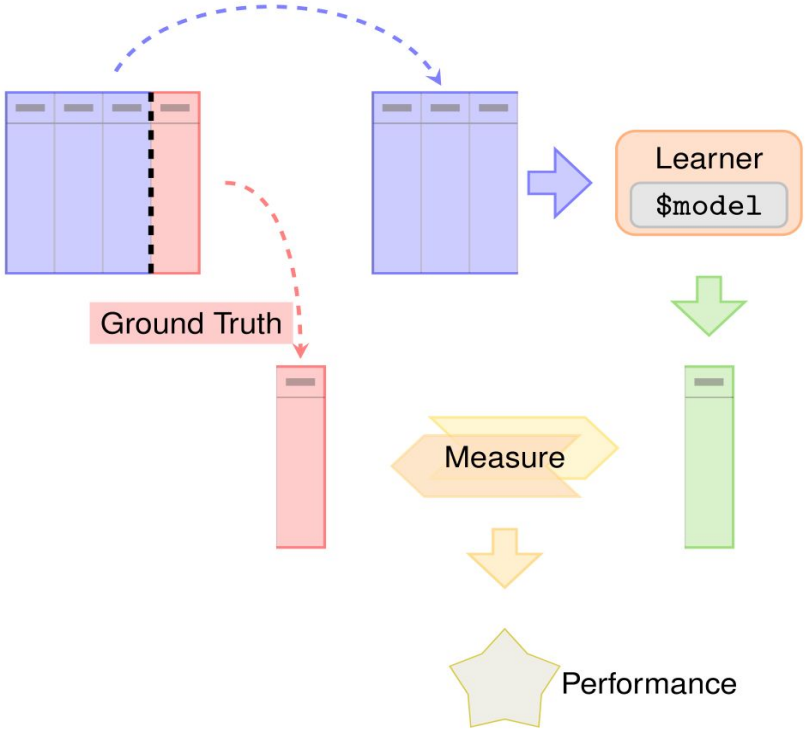


Treinamento e Predição



https://introduction-to-machine-learning.netlify.app/chapter09-01-mlr3-intro

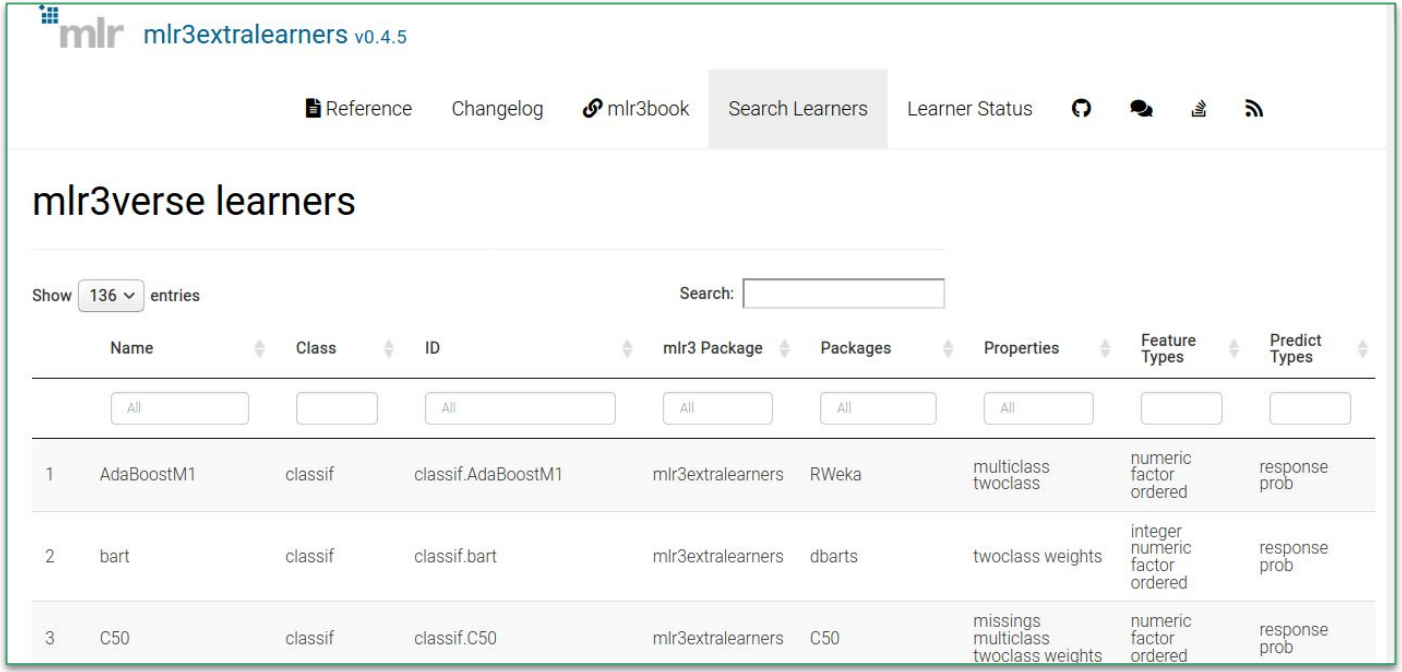
Validação



https://introduction-to-machine-learning.netlify.app/chapter09-01-mlr3-intro

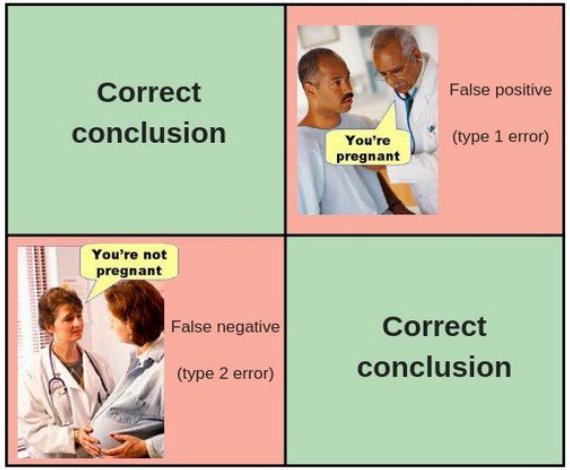
mlr3: Learners

- 136 modelos:
  https://mlr3extralearners.mlr-org.com/articles/learners/list_learners.html



Métricas

Predicted Class \ True Class	Positive	Negative
Positive	TP	FP
Negative	FN	TN



Predicted Class \ True Class	Positive	Negative
Positive	5	50
Negative	10	10000

Accurácia = (TP + TN) / (TP+TN+FP+FN)= (5 + 10000) / (5 + 50 + 10 + 10000)= 0.994 = **99.4%**

Precisão = TP / (TP + FP) = 5 / (5 + 50) = 0.09 = **9%**

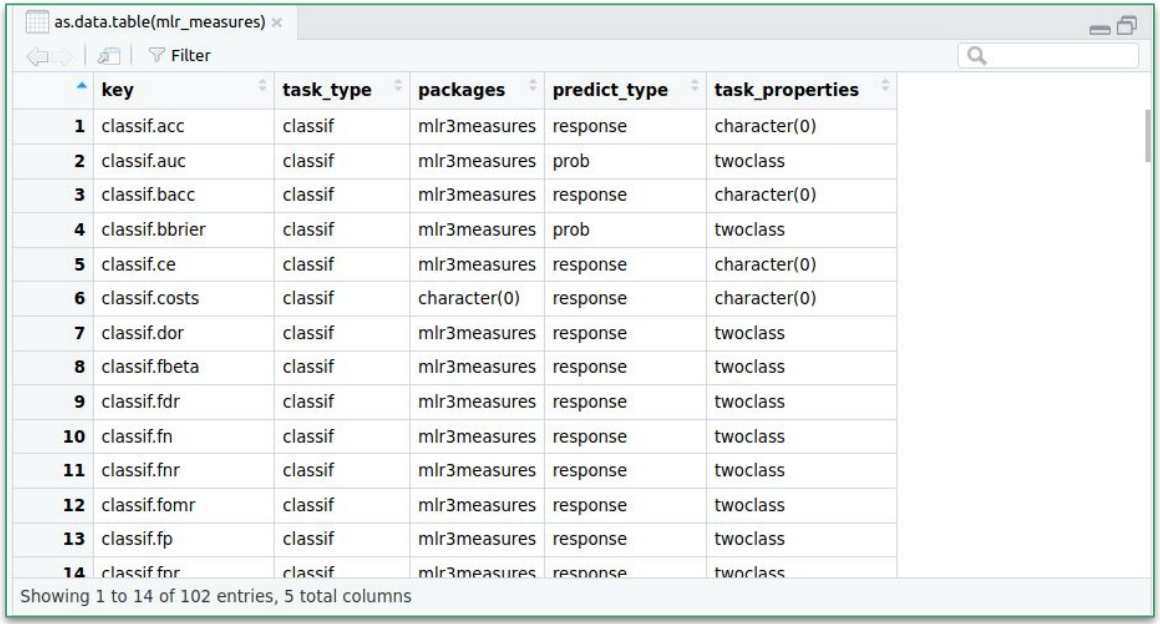
Sensitividade, Recall = TP / (TP + FN) = 5 / (5 + 10) = 0.333 = **33.3%**

Especificidade = TN / (TN + FP) = 10000 / (10000 + 50) = 0.995 = **99.5%**

Acurácia Balanceada = (Sensitividade + Especificidade)/2 = 0.664 = **66.4%**

mlr3: Measures

- 102 métricas:
  https://mlr3.mlr-org.com/reference/mlr_measures.html



as.data.table(mlr_measures)

Filter

	key	task_type	packages	predict_type	task_properties
1	classif.acc	classif	mlr3measures	response	character(0)
2	classif.auc	classif	mlr3measures	prob	twoclass
3	classif.bacc	classif	mlr3measures	response	character(0)
4	classif.bbrier	classif	mlr3measures	prob	twoclass
5	classif.ce	classif	mlr3measures	response	character(0)
6	classif.costs	classif	character(0)	response	character(0)
7	classif.dor	classif	mlr3measures	response	twoclass
8	classif.fbeta	classif	mlr3measures	response	twoclass
9	classif.fdr	classif	mlr3measures	response	twoclass
10	classif.fn	classif	mlr3measures	response	twoclass
11	classif.fnr	classif	mlr3measures	response	twoclass
12	classif.fomr	classif	mlr3measures	response	twoclass
13	classif.fp	classif	mlr3measures	response	twoclass
14	classif.fpr	classif	mlr3measures	response	twoclass

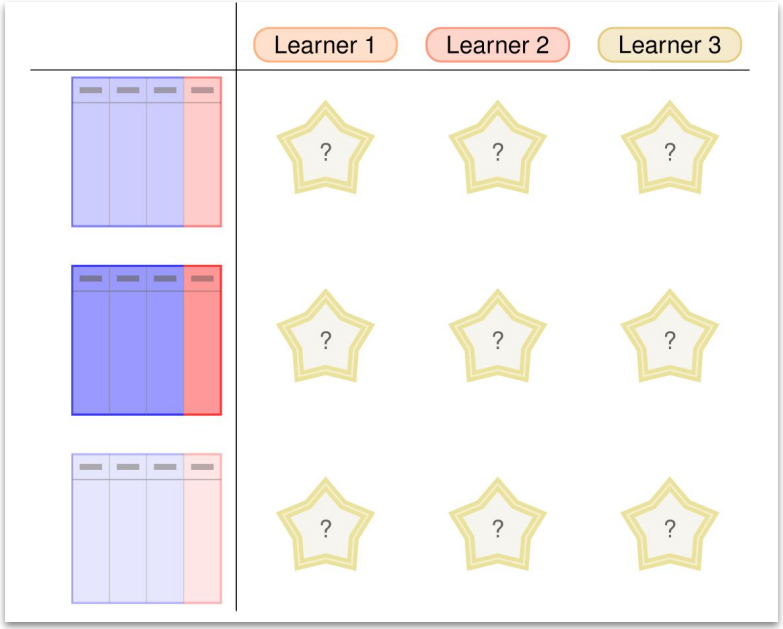
Showing 1 to 14 of 102 entries, 5 total columns

Benchmark

```
library("mlr3learners")
learners = list(lrn("classif.rpart"), lrn("classif.kknn"))
tasks = list(tsk("iris"), tsk("sonar"), tsk("wine"))
```

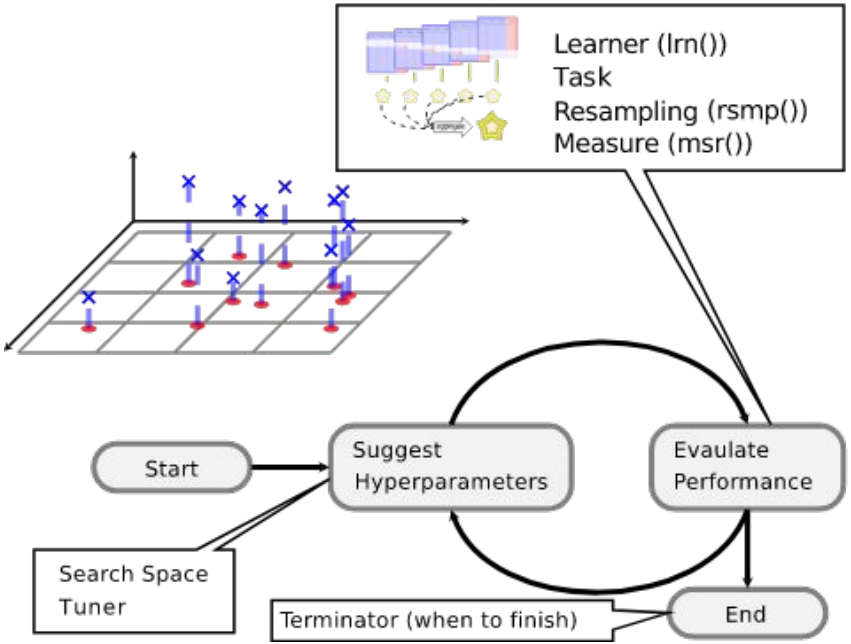
```
design = benchmark_grid(tasks, learners, cv5)
bmr = benchmark(design)
```

```
bmr_ag = bmr$aggregate()
bmr_ag[, c("task_id", "learner_id", "classif.ce")]
#>    task_id    learner_id classif.ce
#> 1:    iris classif.rpart      0.060
#> 2:    iris  classif.kknn      0.060
#> 3:   sonar classif.rpart      0.279
#> 4:   sonar  classif.kknn      0.168
#> 5:    wine classif.rpart      0.101
#> 6:    wine  classif.kknn      0.051
```

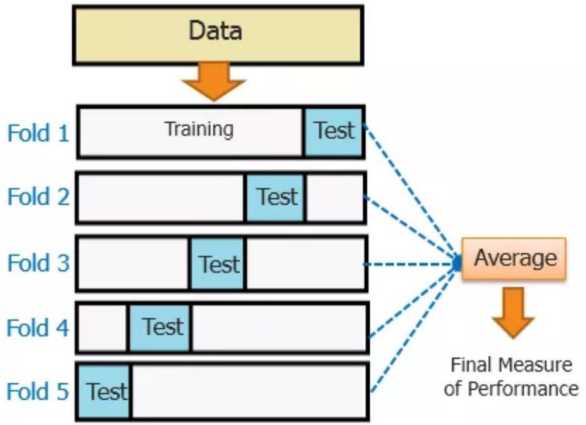


3. Otimização de modelos

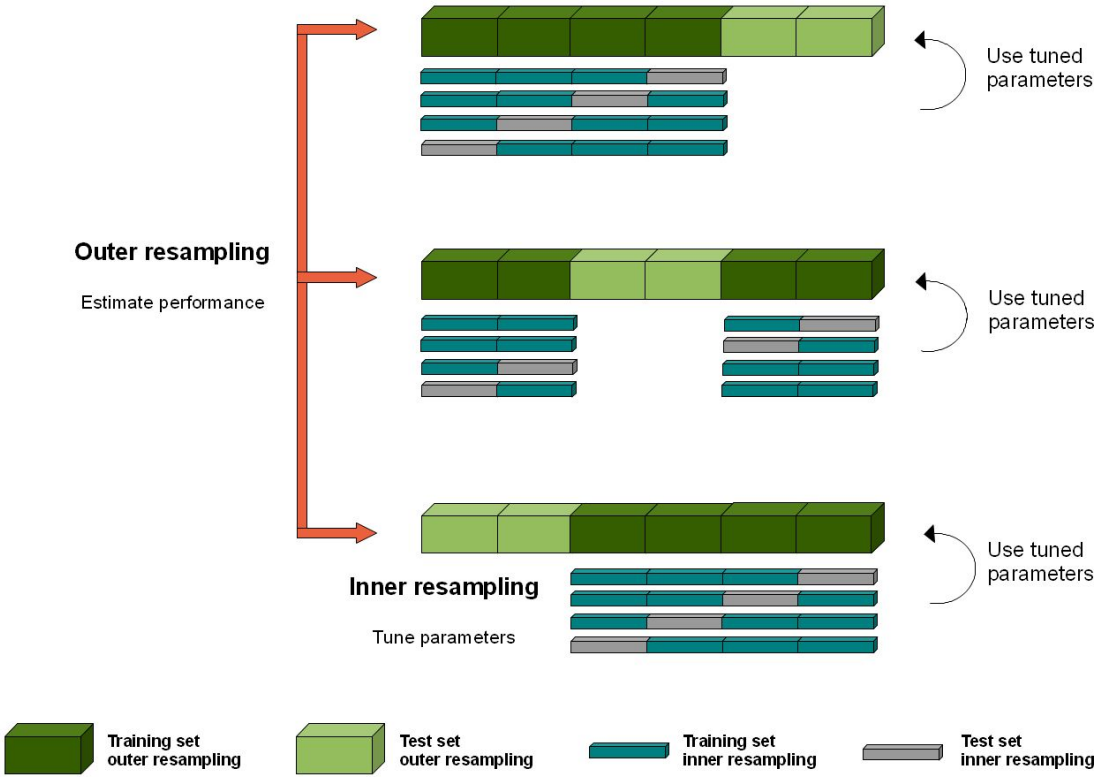
Ajuste de hiperparâmetros



Outer / inner resampling



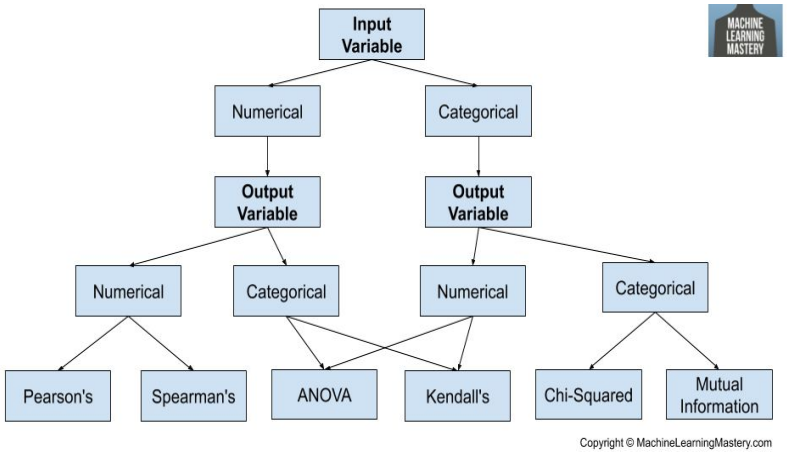
Validação cruzada



Seleção de características

- Filtering: um algoritmo externo computa um ranking das características
  - métodos estatísticos
  - feature importance
  - ganho de informação / entropia
  - independência/não-redundância

  https://mlr3filters.mlr-org.com



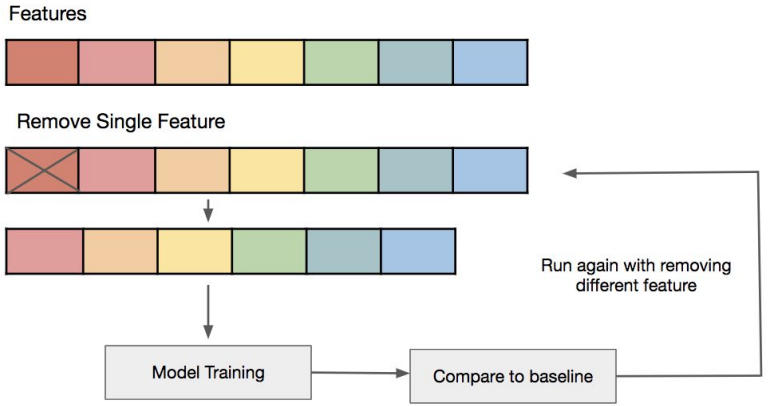
https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Seleção de características

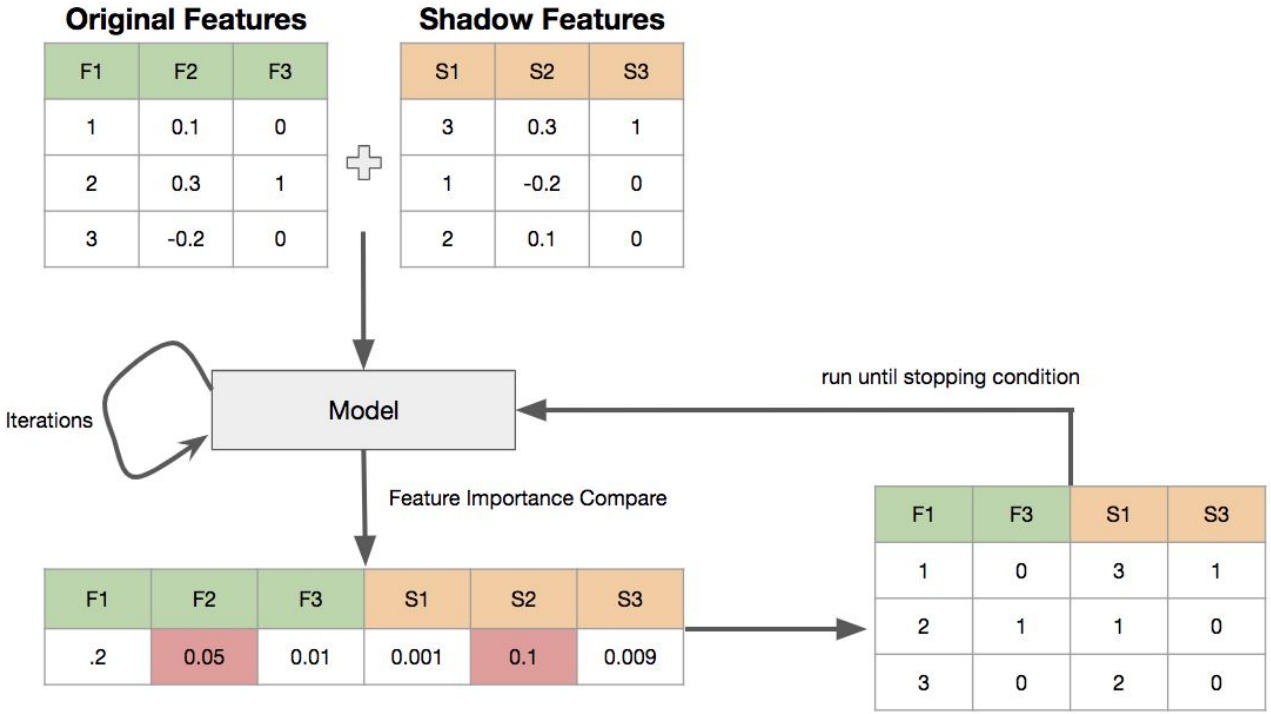
- Métodos Wrapper: um algoritmo de otimização seleciona um subconjunto das características, avalia a performance, e seleciona um novo conjunto.

  - `random_search(batch_size)`
  - `Exhaustive_search(max_features)`
  - `sequential(strategy)`
  - `rfe(feature_fraction, recursive)` Recursive feature elimination
  - `design_points(batch_size, design)` User supplied feature subsets.

  https://mlr3fselect.mlr-org.com



Boruta

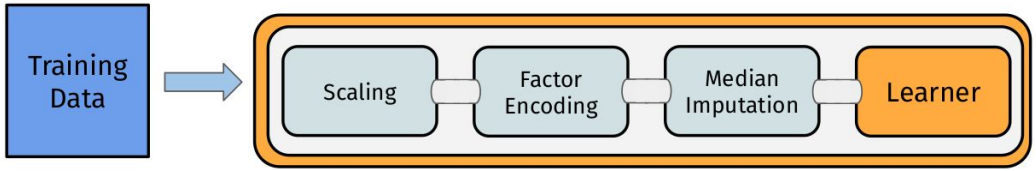


https://cran.r-project.org/web/packages/Boruta/index.html

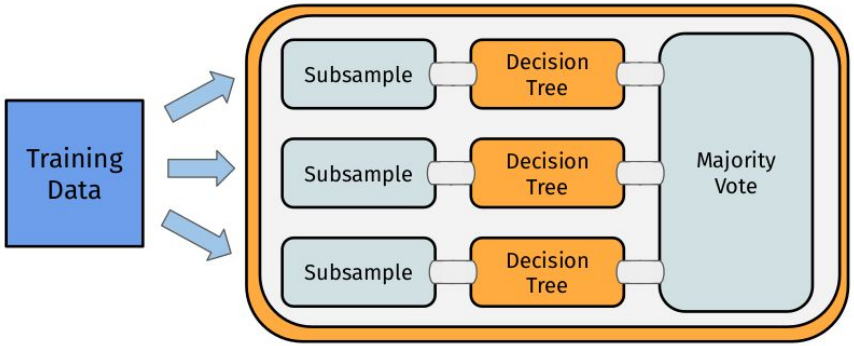
4. Pipelines

Pipelines de Aprendizagem de Máquina

- Pré-processamento



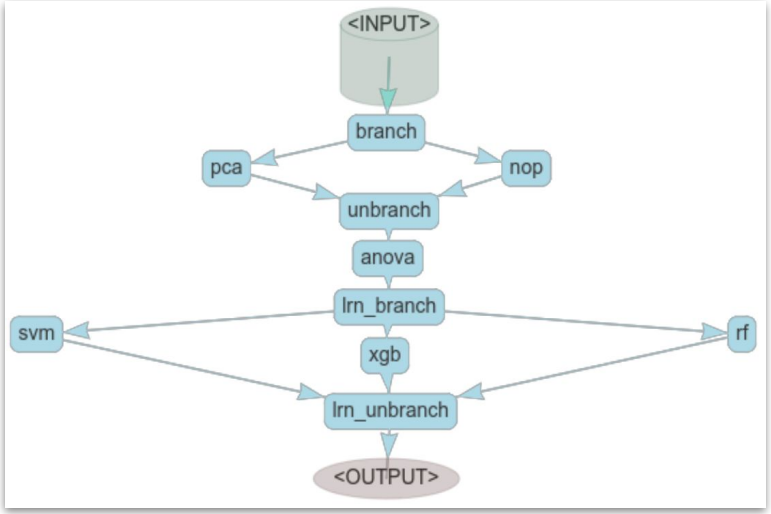
- Comitê de máquinas



https://introduction-to-machine-learning.netlify.app/chapter09-04-mlr3-pipielines

Pipelines de Aprendizagem de Máquina

```
p1 = ppl("branch", list(
  "pca" = po("pca"),
  "nothing" = po("nop")
))
p2 = flt("anova")
p3 = ppl("branch", list(
  "svm" = lrn("classif.svm", id = "svm", kernel = "radial",
    type = "C-classification"),
  "xgb" = lrn("classif.xgboost", id = "xgb"),
  "rf" = lrn("classif.ranger",  id = "rf")
), prefix_branchops = "lrn_")
gr = p1 %>>% p2 %>>% p3
glrn = GraphLearner$new(gr)
```



https://introduction-to-machine-learning.netlify.app/chapter09-04-mlr3-pipelines

5. Hands-on tutorial

https://github.com/terrematte/minicurso_mlr3

Alternativas ao mlr3

- MLSeq: Machine Learning Interface for RNA-Seq Data
  https://www.bioconductor.org/packages/release/bioc/html/MLSeq.html
- Tensorflow
  https://tensorflow.rstudio.com
- Tidymodels
  https://www.tidymodels.org
- Caret: Classification And REgression Training
  http://topepo.github.io/caret/index.html

Bookmarks

- A free, interactive course using tidy tools
  https://supervised-ml-course.netlify.app
- Introduction to machine learning (I2ML)
  https://introduction-to-machine-learning.netlify.app
- R software handbook
  https://bookdown.org/aschmi11/RESMHandbook
- Working with data in the tidyverse
  https://rpubs.com/Sergio_Garcia/working_with_data_in_the_tidyverse
- mlr3 book
  https://mlr3book.mlr-org.com
- mlr3 cheatsheets
  https://cheatsheets.mlr-org.com
- mlr3 gallery
  https://mlr3gallery.mlr-org.com
- Comparing mlr3pipelines to other frameworks
  https://mran.microsoft.com/snapshot/2019-12-22/web/packages/mlr3pipelines/vignettes/comparison_mlr3pipelines_mlr_sklearn.html
- Caret vs. tidymodels - comparing the old and new
  https://konradsemsch.netlify.app/2019/08/caret-vs-tidymodels-comparing-the-old-and-new