TEC | Tecnológico de Costa Rica

Aprendizaje Automático

Tecnológico de Costa Rica

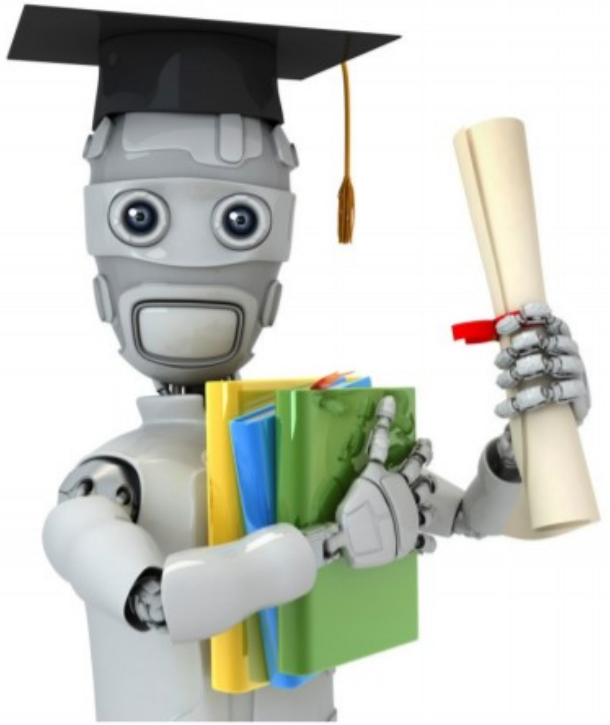
Programa de Ciencia de Datos

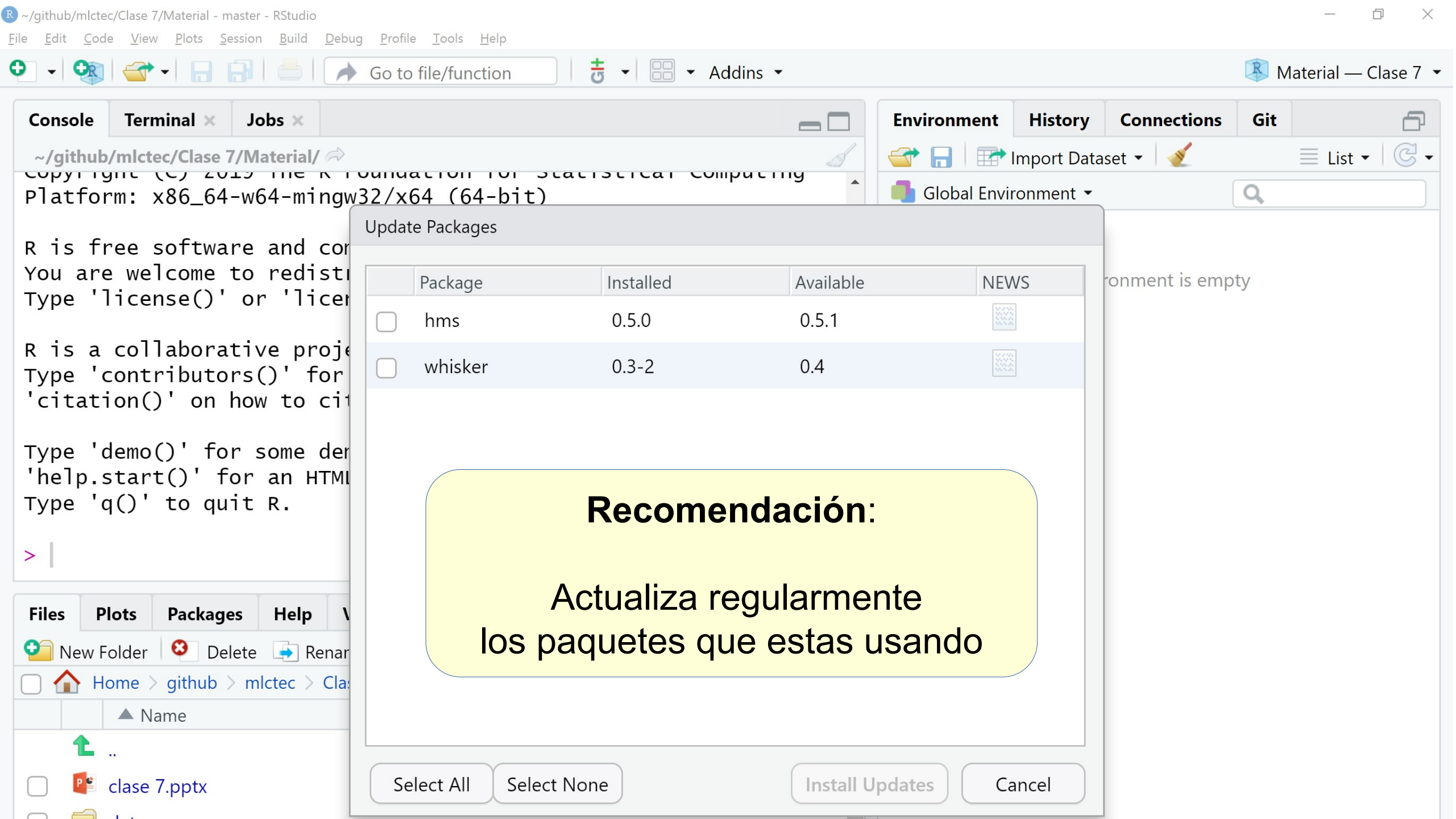
Frans van Dunné

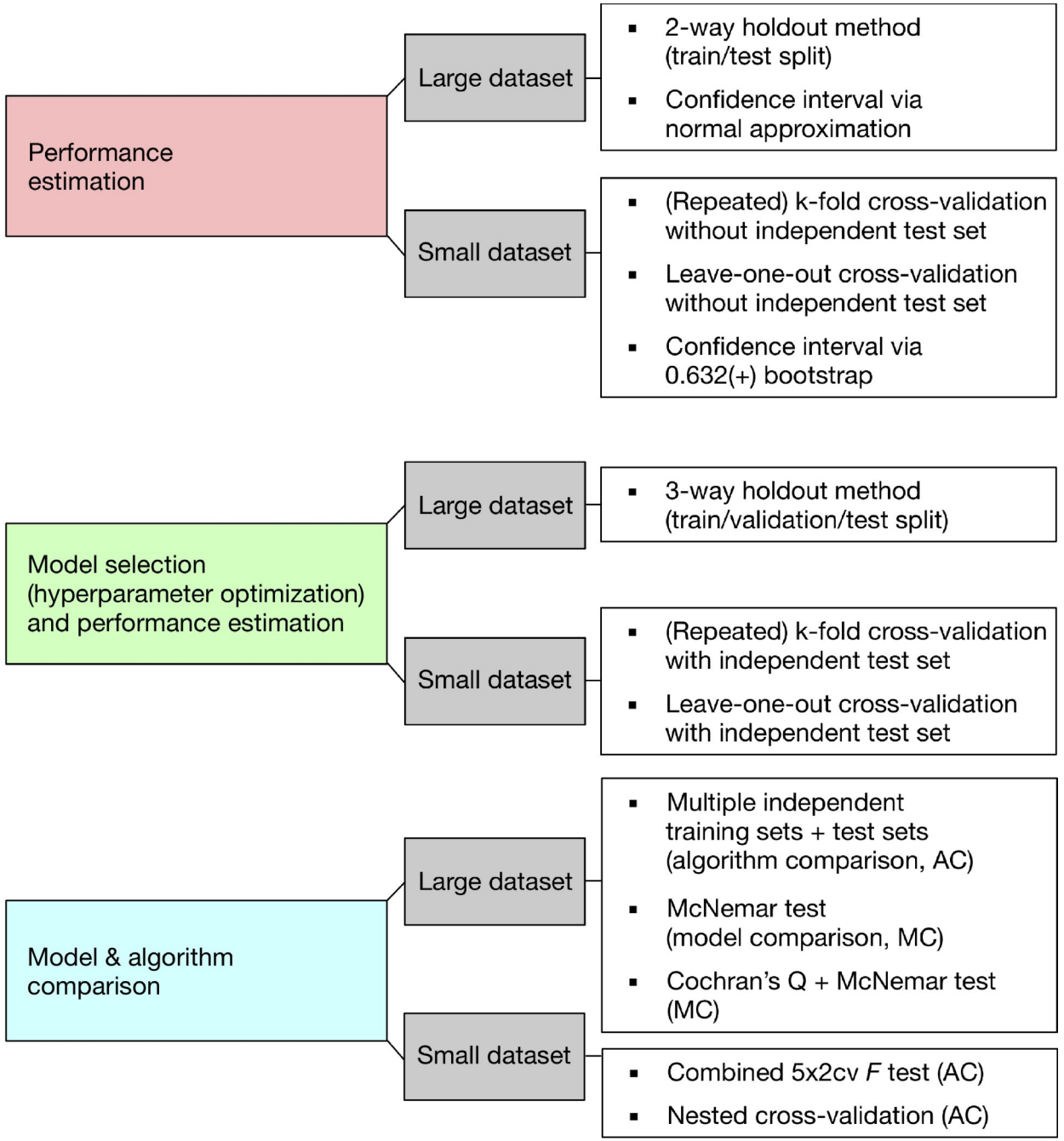
Agenda

TEC | Tecnológico de Costa Rica

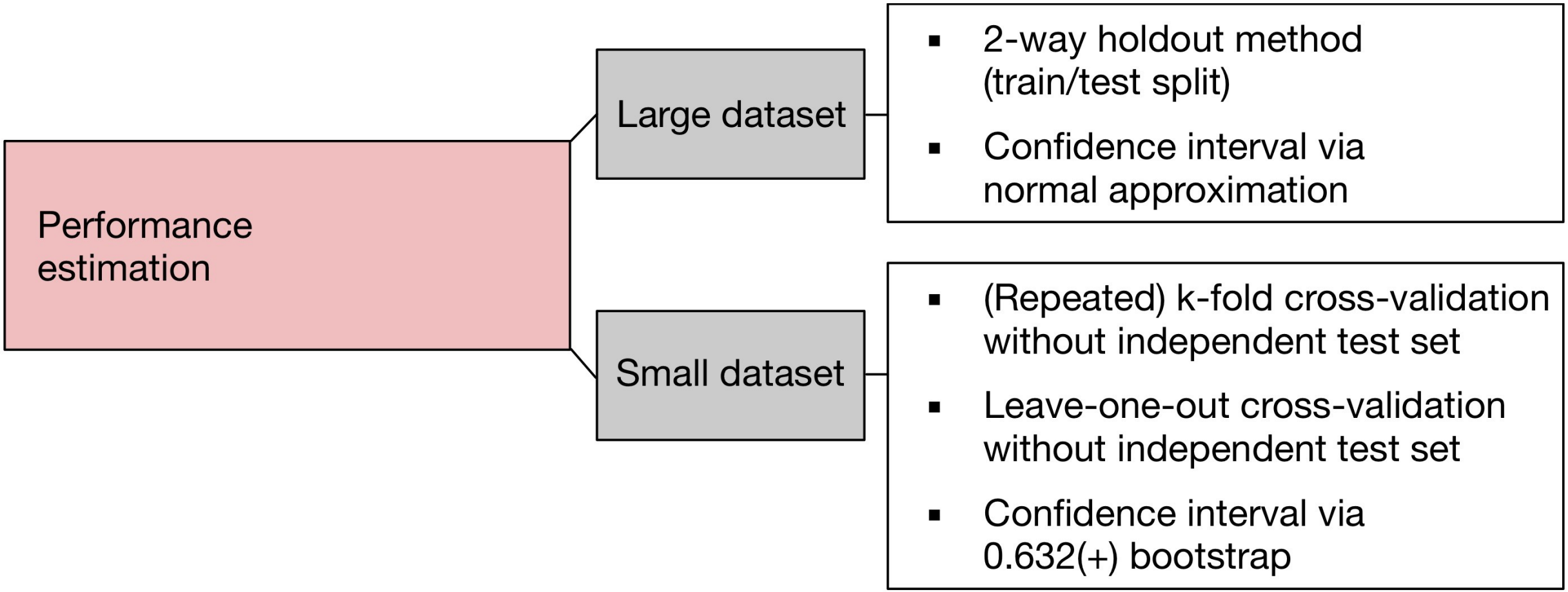
- **Aprendizaje Automático**
  - Validación
    - Matriz de confusion
    - ROC



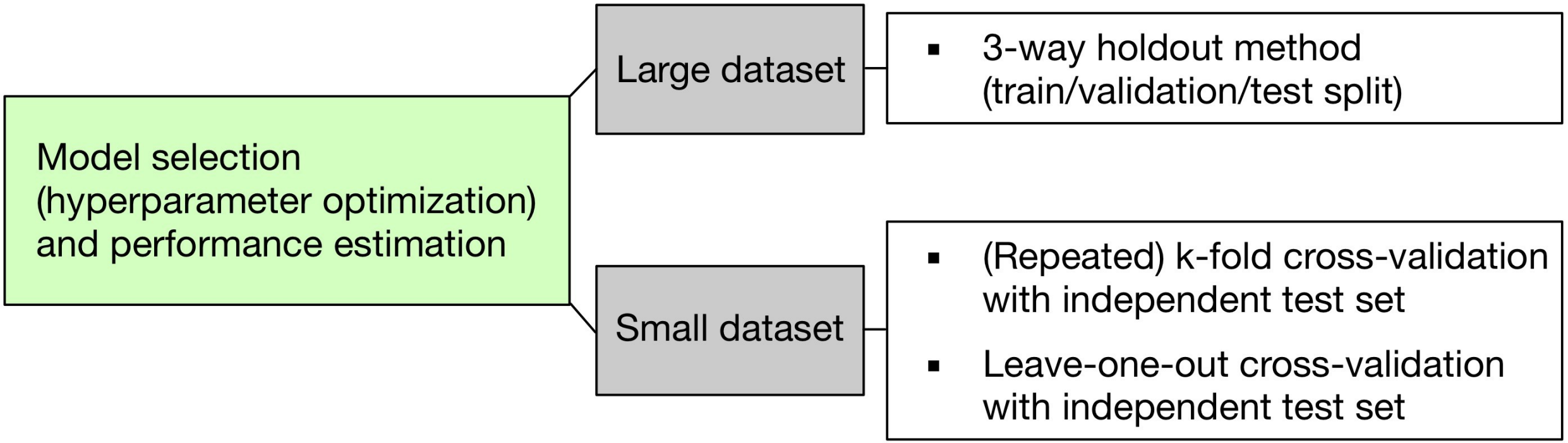

~/github/mlctec/Clase 7/Material - master - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Material — Clase 7
Console
Terminal
Jobs
~/github/mlctec/Clase 7/Material/
Platform: x86_64-w64-mingw32/x64 (64-bit)
R is free software and
You are welcome to
Type 'license()' or
R is a collaborative
Type 'contributors()' for
'citation()' on how to
Type 'demo()' for some
'help.start()' for an
Type 'q()' to quit R.
Environment
History
Connections
Git
Import Dataset
List
Global Environment
Update Packages
Package
Installed
Available
NEWS
hms
0.5.0
0.5.1
whisker
0.3-2
0.4
Recomendación:
Actualiza regularmente
los paquetes que estas usando
Select All
Select None
Install Updates
Cancel
Files
Plots
Packages
Help
New Folder
Delete
Home
github
mlctec
Name
clase 7.pptx



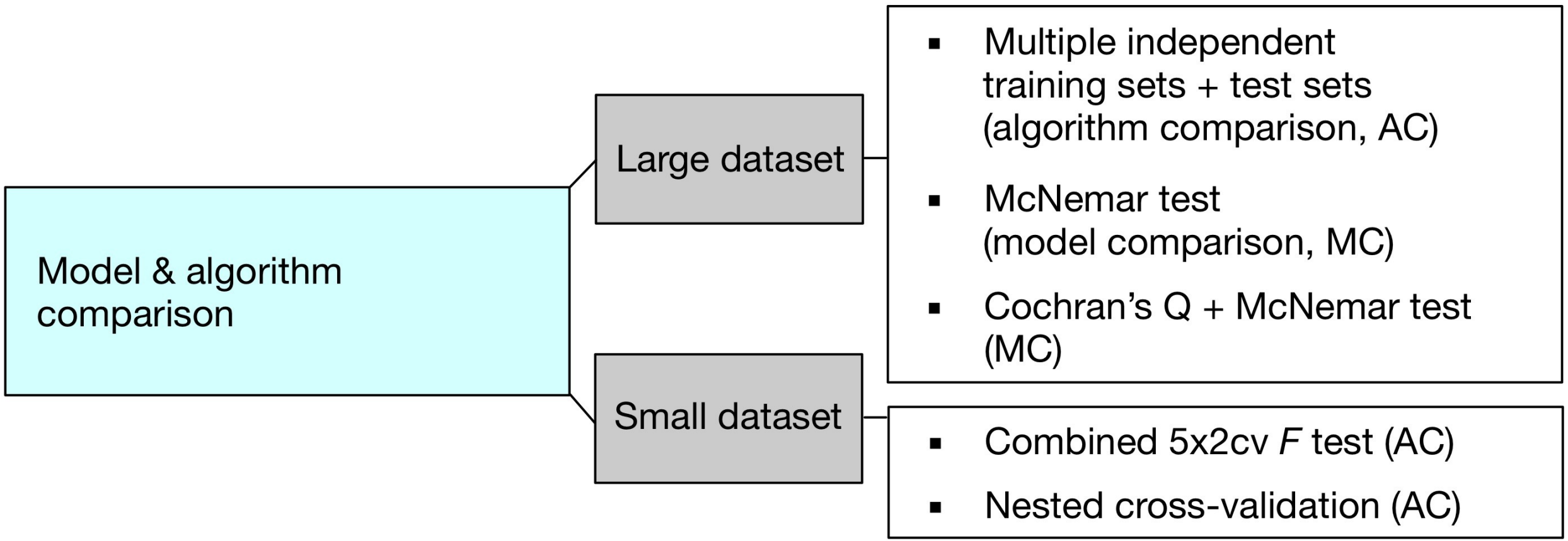
https://sebastianraschka.com/blog/2018/model-evaluation-selection-part4.html



https://sebastianraschka.com/blog/2018/model-evaluation-selection-part4.html



https://sebastianraschka.com/blog/2018/model-evaluation-selection-part4.html



https://sebastianraschka.com/blog/2018/model-evaluation-selection-part4.html

Matriz de confusión

TEC | Tecnológico de Costa Rica

- En el campo de la inteligencia artificial una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.

Matriz de Confusion		**Predicho**			
		Negativo	Positivo		
Real	Negativo	a	b	***Verdadero Negativo (True negative rate)***	a/(a+b)
	Positivo	c	d	***Exactitud***	d/(c+d)
		Sensibilidad	***Especifidad***	**Precisión**=(a+d)/(a+b+c+d)	
		d/(d+c)	a/(a+b)		

Matriz de confusión

TEC | Tecnológico de Costa Rica

- **VP** es la cantidad de *positivos* que fueron *clasificados correctamente* como positivos por el modelo.
- **VN** es la cantidad de *negativos* que fueron *clasificados correctamente* como negativos por el modelo.
- **FN** es la cantidad de *positivos* que fueron *clasificados incorrectamente* como negativos.
- **FP** es la cantidad de *negativos* que fueron *clasificados incorrectamente* como positivos.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Métricas

TEC | Tecnológico de Costa Rica

- Exactitud (Acurracy)
  - En general, que porcentage de la data clasifica correctamente?

$$Exactitud = \frac{VP + VN}{Total}$$

- Tasa de error (Misclassification Rate)
  - En general, que porcentage de la data clasifica incorrectamente?

$$\text{Tasa de error} = \frac{FP + FN}{Total}$$

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Métricas

TEC | Tecnológico de Costa Rica

- Exactitud (Acurracy)
  - En general, que porcentage de la data clasifica correctamente?

$$Exactitud = \frac{VP + VN}{Total}$$

- Tasa de error (Misclassification Rate)
  - En general, que porcentage de la data clasifica incorrectamente?

$$\text{Tasa de error} = \frac{FP + FN}{Total}$$

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Calculenlos!

Metricas

TEC | Tecnológico de Costa Rica

- Sensibilidad, exhaustividad, Tasa de verdaderos positivos
  - Recall
  - Sensitivity
  - True Positive Rate
  - Cuando la clase es positiva, que porcentage logra clasificar?

$$\text{Sensibilidad} = \frac{VP}{\text{Total Positivos}}$$

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

- Especificidad, tasa de verdaderos negativos
  - Especificity
  - True Negative Rate
  - Cuando la clase es negativa, que porcentage logra clasificar?

$$\text{Especificidad} = \frac{VN}{\text{Total Negativos}}$$

Metricas

TEC | Tecnológico de Costa Rica

- Sensibilidad, exhaustividad, Tasa de verdaderos positivos
  - Recall
  - Sensitivity
  - True Positive Rate
  - Cuando la clase es positiva, que porcentage logra clasificar?

$$\text{Sensibilidad} = \frac{VP}{\text{Total Positivos}}$$

- Especificidad, tasa de verdaderos negativos
  - Especificity
  - True Negative Rate
  - Cuando la clase es negativa, que porcentage logra clasificar?

$$\text{Especificidad} = \frac{VN}{\text{Total Negativos}}$$

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Calculenlos!

Métricas

TEC | Tecnológico de Costa Rica

- Precisión
  - Cuando predice positvos, que porcentage clasifica correctamente?

$$\text{Precisión} = \frac{VP}{\text{Total clasificados positivos}}$$

- Valor de predicción negativo
  - Cuando predice negativo, que porcentage clasifica correctamente?

$$\text{VPN} = \frac{VN}{\text{Total clasificados negativos}}$$

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Métricas

TEC | Tecnológico de Costa Rica

- Precisión
  - Cuando predice positvos, que porcentage clasifica correctamente?

$$\text{Precisión} = \frac{VP}{\text{Total clasificados positivos}}$$

- Valor de predicción negativo
  - Cuando predice negativo, que porcentage clasifica correctamente?

$$\text{VPN} = \frac{VN}{\text{Total clasificados negativos}}$$

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

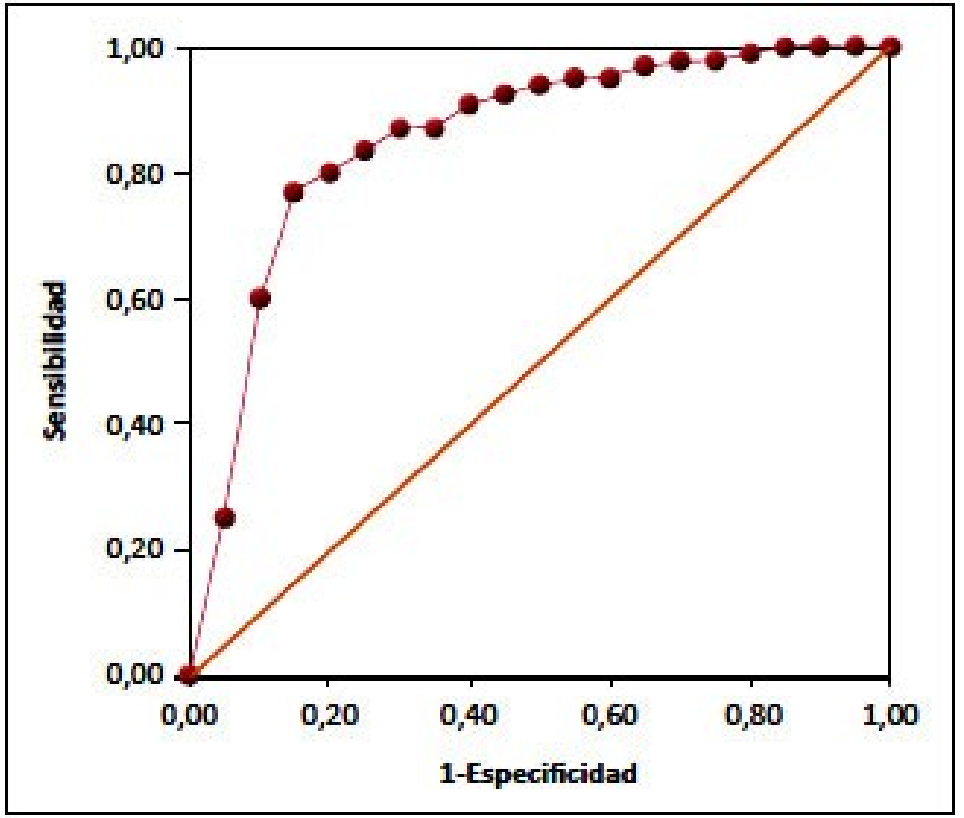
Calculenlos!

Curva ROC

TEC | Tecnológico de Costa Rica

- Curva ROC (acrónimo de Receiver Operating Characteristic, o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación.

- El análisis de la curva ROC, o simplemente análisis ROC, proporciona herramientas para seleccionar los modelos posiblemente óptimos y descartar modelos subóptimos independientemente de (y antes de especificar) el coste de la distribución de las dos clases sobre las que se decide. La curva ROC es también independiente de la distribución de las clases en la población (en diagnóstico, la prevalencia de una enfermedad en la población). El análisis ROC se relaciona de forma directa y natural con el análisis de coste/ beneficio en toma de decisiones diagnósticas.



Some R Packages for ROC Curves

2019-03-01

by Joseph Rickert

In a recent post, I presented some of the theory underlying ROC curves, and outlined the history leading up to their present popularity for characterizing the performance of machine learning models. In this post, I describe how to search CRAN for packages to plot ROC curves, and highlight six useful packages.

Although I began with a few ideas about packages that I wanted to talk about, like ROCR and pROC, which I have found useful in the past, I decided to use Gábor Csárdi's relatively new package pkgsearch to search through CRAN and see what's out there. The `package_search()` function takes a text string as input and uses basic text mining techniques to search all of CRAN. The algorithm searches through package text fields, and produces a score for each package it finds that is weighted by the number of reverse dependencies and downloads.

```
library(tidyverse)  # for data manipulation
library(dlstats)    # for package download stats
library(pkgsearch)  # for searching packages
```

After some trial and error, I settled on the following query, which includes a number of interesting ROC-related packages.


```
rocPkgShort <- rocPkg %>%
               filter(maintainer_name != "ORPHANED", score > 190) %>%
               select(score, package, downloads_last_month) %>%
               arrange(desc(downloads_last_month))
head(rocPkgShort)
## # A tibble: 6 x 3
##   score package  downloads_last_month
##   <dbl> <chr>                   <int>
```

Ejercicio

TEC | Tecnológico de Costa Rica

- Desarrolle el siguiente ejercicio
  - Traspase el siguiente link un proyecto en Rstudio.
  -
  - https://rpubs.com/chzelada/275494
  - Ahora reemplaza la creacion del ROC con una de las funciones de los paquetes anteriores.

Evaluación de modelos de clasificación

Carlos Zelada

10/5/2017

```
library(ISLR)
library(MASS)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

Selección de modelos

A

		Model 2 correct	Model 2 wrong
Model 1	correct	9959	11
Model 1	wrong	1	29

B

		Model 2 correct	Model 2 wrong
Model 1	correct	9945	25
Model 1	wrong	15	15

`mcnemar.test {stats}`

https://sebastianraschka.com/blog/2018/model-evaluation-selection-part4.html

Bibliografía

TEC | Tecnológico de Costa Rica

https://sebastianraschka.com/blog/2018/model-evaluation-selection-part4.html

https://www.researchgate.net/figure/Figura-5-Representacion-grafica-de-curvas-ROC-de-dos-pruebas-diagnosticas-hipoteticas-A_fig3_323485977

https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0716-10182012000200003

https://es.wikipedia.org/wiki/Curva_ROC