

# Fundamentos de la Programación Estadística y Data Science en R - Versión tidyverse

Un curso de R orientado a estudiantes de ciencias duras, sociales, profesionales, técnicos, etc.

Docente: Germán Rosati

## Presentación y objetivo del curso:

Debido a su carácter de software libre y a la creciente comunidad de usuarios el lenguaje R se ha convertido en algo así como la lingua franca dentro del análisis estadístico. El presente seminario se propone realizar una introducción a algunos conceptos fundamentales de la programación estadística en R. A su vez, se hará énfasis en la implementación de análisis estadísticos básicos (descriptivos y regresiones) en R. A su vez, el curso presentará algunos elementos teóricos de la minería de datos/aprendizaje automático (balance sesgo-variancia, overfitting, etc.) y revisará algunos algoritmos para la estimación de árboles (ID4, C4.5, CART y random forest).

El curso se propone que los alumnos:

- se familiaricen con aspectos relevantes de la programación estadística en lenguaje R y con el llamado **tidyverse** en particular
- logren implementar e interpretar análisis estadísticos descriptivos y modelos de regresión en lenguaje R
- incorporen algunos conceptos fundamentales del data mining/aprendizaje automático,
- conozcan generalidades de algunos algoritmos para la generación de árboles de decisión (ID4, C4.5, CART y random forest) y su implementación en lenguaje R,
- logren identificar situaciones de aplicación de este tipo de modelos a problemas de investigación básica y aplicada

## Requisitos para la cursada y aprobación

Conocimientos básicos de estadística descriptiva y cierta familiaridad con el modelo de regresión lineal y logística. Será útil (pero no absolutamente necesario) alguna experiencia en programación estadística (sea en SPSS, Stata o similar) Para la aprobación del curso se requiere:

1. un mínimo de asistencia del 80% sobre el total de clases y
2. la entrega y aprobación de una monografía final

## Fuentes

El material para el curso fue extraído y transformado de diversas fuentes.

- Curso R Programming - Coursera
- Curso Programación Estadística en R - Coursera
- Materiales didácticos de Introduction to Statistical Learning, escrito por James, Witten, Hastie y Tibshirani
- Materiales de Kelly Black
- R for Data Science
- Ciencia de Datos para Gente Sociable
- Quick R Tutorial
- R Tutorial
- R Cheat-Sheet
- R Reference Card
- A very quick introduction to ggplot2

## Contenidos resumidos

- *Unidad 1a. Elementos de programación estadística en R:* Objetos en R (vectores, matrices, data frames y listas). Introducción al **tidyverse**: data wrangling (`select()`, `filter()`, `arrange()`, `mutate()`, `summarise()`, `group_by()`, `left_join()`). Estructuras de control: `for`, `if`. Uso e implementación de funciones ad-hoc. Importación y exportación de datos (`.csv`, `.txt`, `.tab`, `.sav`, etc.).
- *Unidad 1b. Visualización y generación de gráficos en R:* Nociones de graficación (forma, color, tamaño, color). Niveles de medición y gráficos adecuados. Introducción a **ggplot2**: `ggplot()`, `geom_points()`, `geom_smooth()`, `aes()`, `facet_wrap()`, `facet_grid()`.
- *Unidad 2. Nociones básicas de data mining/aprendizaje automático:* Tipos de problemas en aprendizaje supervisado: clasificación y regresión. Error de entrenamiento (training error), error de prueba (test error). Sobre-ajuste. Balance entre el sesgo y la variancia de un modelo. Métodos de estimación del error: partición del dataset, validación cruzada. Aplicaciones en R.
- *Unidad 3a. Introducción a los problemas de regresión y clasificación en R:* Implementación y análisis de modelos de regresión lineal y logística. Evaluación del modelo: supuestos, ajuste, estimación de error de generalización. Extensiones del modelo lineal y logístico: variables cualitativas, no linealidad, etc. Funciones `lm`, `glm` y `predict`. Funciones `lm()`, `glm()` y `predict()`.
- *Unidad 4. Clasificadores basados en árboles:* Generalidades. Algoritmos ID4, C4.5 y CART. Partición múltiple y binaria, medidas de pureza de nodos. Crecimiento (growing) y podado (prunning) de árboles de decisión. Balance entre costo y complejidad del árbol. Aplicaciones en R (paquetes `tree`, `rpart` y `randomForest`).
- *-Opcional- Unidad 5. Introducción al análisis de secuencias con el paquete TraMineR en R:* Estructuras de datos para el análisis de secuencia:

inspección gráfica, análisis descriptivo (largo de secuencias, tiempo de estados, tasas de transición, etc.), medidas de disimilaridad entre secuencias, identificación de secuencias relevantes, tipologías de secuencias (clustering, etc.).

## Bibliografía básica

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications in R*, Berlin: Springer.
- R Core Development Team, (2000), *Introducción a R. Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos*.
- Teter, P. (2011), *R Cookbook. Proven recipes for data analysis, statistics and graphics*, New York: O Reilly.