UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad

www.upv.es

bigdata.inf.upv.es

Herramientas estadísticas para Big Data

Máster **Big Data** Analytics

Valencia, Octubre 2016

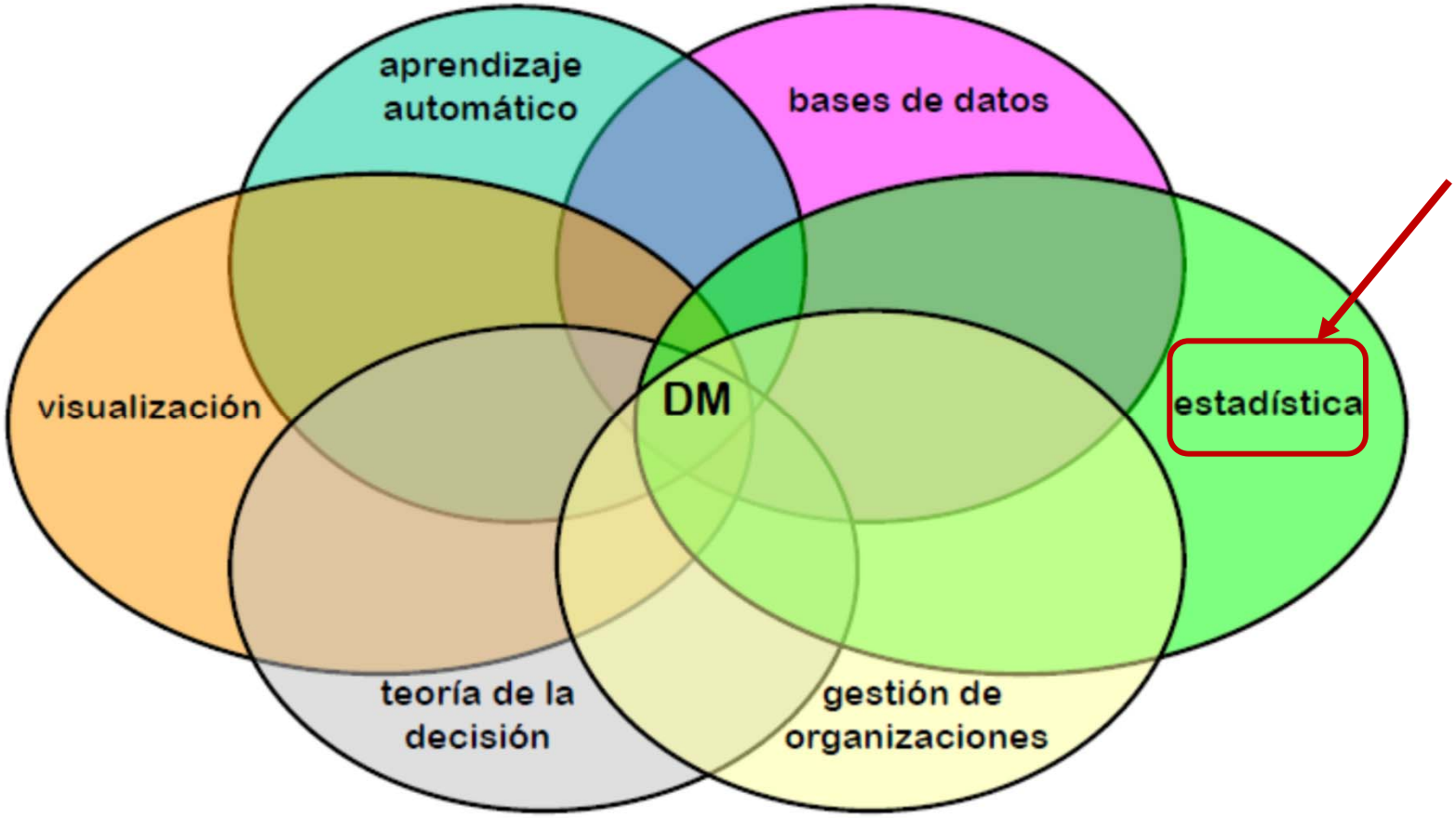
Mónica Clemente | Ana Debón | Elena Vázquez

Profesores

- Mónica Clemente Císcar
- Ana Debón Aucejo
- Elena Vázquez Barrachina

Departamento de Estadística e Investigación Operativa Aplicadas y Calidad - UPV

¿Por qué necesitamos la Estadística?



Probar teorías cuantitativamente

Software

- **R**

http://www.r-project.org/

- **R Studio**

http://www.rstudio.com/

Contenidos parte 1

Parte 1: 2 sesiones de 5 h - viernes 7 y viernes 14

1. Conceptos básicos
2. Probabilidad y variables aleatorias
3. Inferencia en muestras grandes
4. Técnicas de muestreo
5. Preprocesamiento de datos

Parte 2: 2 sesiones de 5 h

6. Técnicas de clasificación supervisada **sábado 8**
7. Evaluación de los modelos de clasificación **sábado 15**

Contenidos Viernes 13

1. CART
   - Clasificación
   - Regresión
2. Evaluación de clasificadores
3. Evaluación de modelos de regresión

Contenidos Sábado 14

1. Clasificadores individuales
   - Árboles de decisión
   - Regresión Logística
2. Métodos de combinación de clasificadores:
   - Random Forest

Evaluación

- **Parte 1**: Prof. Elena Vázquez
  - Una **trabajo** relativo a **Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos** (primera y última sesiones) enviado a través de la herramienta **Tareas de PoliformaT**

- **Parte 2**: Prof. Mónica Clemente y Prof. Ana Debón

Revisión de los contenidos

Parte 2: Prof. Mónica Clemente y Prof. Ana Debón

- Trabajo no presencial sesión 3 y 4 (viernes 13 y sábado 14): Leer el documento de resultados de la liga de futbol que hay en recursos y responder a las preguntas cortas.

- Realización del examen que se encuentra en el Poliformat "Evaluación Resultados_partidos_futbol"
  - Estará activo durante un mes, a partir del lunes 16 de Octubre
  - Se podrá enviar 2 veces y se guardará la mejor nota

Evaluación Parte 2

Parte 2: Prof. Mónica Clemente y Prof. Ana Debón

- Trabajo no presencial sesión 3 y 4 (viernes 13 y sábado 14): Ejecutar el script “TrabajoSesion3y4_Alumnos.R”, siguiendo las indicaciones del fichero “HerramientasEstBigData_TrabajoNoPres_Sesion3y4.pdf”

- Realización del examen que se encuentra en el Poliformat “Evaluación Contenidos-Modelos de clasificación”
  - Estará activo durante un mes, a partir del lunes 16 de Octubre
  - Se podrá enviar 2 veces y se guardará la mejor nota

Paquetes de R *library*

- e1070
- MASS
- pwr
- (Hmisc)
- car
- psych
- corpcor
- GPArotation
- (nFactors)
- (FactoMineR)
- XML
- rvest
- Rcurl
- ggplot2
- rpart
- rpart.plot
- c50
- randomForest
- ROCR

Bibliografía general

Andy Field, Jeremy Miles, and Z.F., 2012. *Discovering statistics using R*, SAGE Publications.

C. Vercellis , 2009. *Business intelligence: data mining and optimization for decision making*. Wiley Online Library.

Diez, D.M., Barr, C.D. & Cetinkaya, M., 2010. *OpenIntro : Statistics Preliminary Edition*, OpenIntro. Available at: https://www.openintro.org/download.php?file=os0&referrer=/stat/textbook.php.

De Jonge, E. & van der Loo, M., 2013. *An introduction to data cleaning with R*, Statistics Netherlands. Available at: http://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf.

Gómez, A. A., 2008. *Estadística básica con R y R-Commander*. Servicio Publicaciones UCA

P. Giudici, 2003. *Applied Data Mining: Statistical Methods for Business and Industry*. Wiley.

Hastie T, Tibshirani R, Friedman J., 2001. The elements of statistical learning. Data mining, inference, and prediction. New York: Springer. (http://knuth.uca.es/moodle/mod/url/view.php?id=1126)

J. Hernández, M. J. Ramírez, and C. Ferri, , 2004. *Introducción a la Minería de Datos*, volume 17. Prentice.

Kuhn, M., & Johnson, K., 2013. *Applied predictive modeling*. New York: Springer.

L. Torgo, 2010. Data mining with R: learning with case studies. Chapman & Hall/CRC.

Sáez Castillo, A.J., 2010. *Métodos Estadísticos con R y R Commander*, Jaén: Universidad de Jaén. Available at: http://cran.r-project.org/doc/contrib/Saez-Castillo-RRCmdrv21.pdf.

Sanchez De Rivera, D.P., 1993. *Estadistica modelos y metodos 1. fundamentos*, Alianza.

Scheaffer, R., Mendenhall, W. & Ott, L., 2007. *Elementos de muestreo*, Editorial Paraninfo.

Enlaces de interés

- **Quick-R** http://www.statmethods.net/
- **R-bloggers** https://www.r-bloggers.com/
- Stack**overflow** http://stackoverflow.com/
- Stack**overflow** en español http://es.stackoverflow.com/
- https://www.coursera.org/course/statistics
- http://spark.rstudio.com/minebocek/dist_calc/

Gracias por vuestra atención

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

www.upv.es