UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Data Science (CDA)

33420 - Ciència de Dades

2023-2024

- José Hernández Orallo, DSIC-UPV, jorallo@upv.es
- Cèsar Ferri Ramírez, DSIC-UPV, cferri@dsic.upv.es
- Behzad Mehrbakhsh, ValGRAI, VRAIN-UPV, bmehrba@upv.es

Description

- Credits: 6.0 (1.5: theory, 3: seminar, 1.5: lab)
  - Theory and seminar will be intertwined.
- Lecturers
  - José Hernández Orallo (jorallo@upv.es)
    - Office 236, 2nd floor DSIC (Bldg. 1F).
    - Attention/tutoring hours: on demand by email.
  - Cèsar Ferri Ramírez (cferri@dsic.upv.es)
    - Office 235, 2nd floor DSIC (Bldg. 1F).
    - Attention/tutoring hours: on demand by email.
  - Behzad Mehrbakhsh(bmehrba@upv.es)
    - Office – lab 203, 2nd floor DSIC (Bldg. 1F).
    - Attention/tutoring hours: on demand by email.

Learning Outcomes

After completion of the course, the student will be able to understand the role of the data scientist in organisations, identify problems and opportunities and deploy solutions using off-the-shelf tools.

- Goals:
  1. recognise the value of data and the business opportunities for the development of data-driven products.
  2. determine the technologies that are needed to handle data efficiently in different environments, different sizes and formats, in order to ease data understanding and analysis.
  3. estimate the complexity and resources that are needed for a data analysis project and establish the measures of cost and success.

Learning Outcomes

- Specific objectives:
  - Realise the value of data and data-driven products.
  - Know the process of converting data into knowledge.
  - Use tools to integrate, prepare and visualise data.
  - Use a data-analysis language or tool to obtain models.
  - Evaluate models.
  - Deploy and exploit knowledge.

Contents

- **Unit 1**: Introduction (4,5h)
  - Data science: the role of the data scientist.
  - The value of the data: examples.
  - The D2K (Data to Knowledge) process.
  - Big Data: challenges and solutions.
- **Unit 2**: Data integration and manipulation (15h)
  - Source types and data repositories.
  - Data gathering, integration and cleansing
  - Data property, privacy and security.
  - Data visualisation and comprehension.
- **Unit 3**: Data analysis (17h)
  - Predictive and descriptive tasks
  - Supervised techniques
  - Non-supervised techniques
  - Model evaluation.
- **Unit 4**: Knowledge exploitation (5,5h)
  - Assistants, prescriptors and recommenders
  - Integration into decision making, dashboards and monitoring.

PLUS:
Introduction to **R** (5,5h)
Introduction to **Py**thon (3,5h)
Project Feedback, pre-**Ev**aluation (4,5h)
Final **Ev**aluation (4,5h)

Evaluation

- Mostly practical evaluation:
  - Short questionnaires **in the classroom** (2): Q1, Q2 (10% each)
  - Short practical assignments (3): L1, L2, L3 (10% each)
    - Portfolio **delivered** on Poliforma't (assignment for each of the 10 practicals). Can be done in couples, but **evaluated individually through interview** at most 3 weeks after the start of that practical.
  - Freelance data scientist project: G1 (50%) * C1 (0-1)
    - Groups of **three students**.
    - Develop the idea of a new product from the use of data (open data, Internet, repositories, etc.) or that could improve an existing procedure with data-acquired knowledge.
    - **Oral presentation (pre-evaluation and final evaluation weeks)**.
    - **Evaluation rubric (G1)**: data value, alternatives and innovation, technical tool integration, project effort and exposition quality.
    - **Co-evaluation rubric (C1)**: percentage of contribution, disposition
    - Presentation **delivered** on Poliforma't.

Course plan

- Mon: 10:00:11:00 (pre-recorded), Tue: 16:30-18:00 (Physical, Room 1G 0.2), Thu: 15:00-17:00 (Physical, Room 1G 0.2)

MON	TUE	Teacher	THU	Teacher	Theory	Seminar/Practicals	Lab block	Deadline	Assessments
Sep-11	Sep-12	Jose	Sep-14	Jose	Pres	Practical1-IntroR			
Sep-18	Sep-19	Jose	Sep-21	Jose	Unit1	Practical2-WorkingWithData (R)	L1		
Sep-25	Sep-26	Jose	Sep-28	Jose	Unit1	Practical3-ggplot (R)	L1		
Oct-02	Oct-03	Jose	Oct-05	Jose	Unit2	Catching up with practicals, starting with the project, Python			
Oct-09	Oct-10	Jose	Bank Holiday	-	Unit2	Practical5-classification (R or python)	L2	Prac2	
Oct-16	Oct-17	Cesar	Oct-19	Jose	Unit2	Practical6-regression (R or python)	L2	Prac3	
Oct-23	Oct-24	Behzad	Oct-26	Jose	Unit3	Practical7-evaluation (R or python)	L3		Q1 - Oct-26
Oct-30	Oct-31	Jose	Nov-02	Jose	Unit3	Practical9-clustering (Python)	L3	Prac5	
Nov-06	Nov-07	Behzad	Nov-09	Behzad	Unit3	Practical10 (Python)	Opt	Prac6	
Nov-13	Nov-14	Behzad	Nov-16	Behzad	Unit3	*Working on the project*		Prac7	
Nov-20	Nov-21	Cesar	Nov-23	Cesar	Unit4	*Working on the project (P10 is optional + 0.5p extra)*		Prac9	
Nov-27	Nov-28	Jose/Cesa	Nov-30	Jose	PRESENTATIONS (PREVALUATION)			Prac10	Q2 - Nov-30
Dec-04	This is a Wed	-	No teaching	-					
Dec-11	Dec-12	-	Dec-14	-	*Working on project feedback for those taking resit*				
Dec-18	Dec-19	Jose	Dec-21	Jose	PRESENTATIONS (RESITS) + Pending evaluations				
January					*COURSE IS OVER: nothing here*				

Days in black mean regular class and in grey mean NO CLASS. Days in magenta mean presentation class.

Honesty Rules

- **Honesty** rules: (hint: anything that makes *you* more productive in the workplace is allowed)
  - Use of ChatGPT, Copilot, GPT4 and other programming/writing/presentation assistants:
    - Not only encouraged but REQUIRED!!! We're in the mechanocene!
    - Free (no-cost) assistants preferable, to avoid disadvantages.
    - **For each deliverable, you have to provide the solution given by the assistant and compare it with yours. Your solution can be based on the assistant's solution, or if the solution from the assistant is perfect, explain why it is so.**
    - Even for the questions, you can ask the assistant, but then discuss the answer, and provide **your* final answer*.
    - **Also use it for the project. But be open about its use and careful (they confabulate, and fabricate things, aka "hallucinate")**
  - Please declare the use of any other resource (especially human experts, etc.).
    - Asking other humans is okay, provided they don't do your work (this goes against *your* productivity).

Recommended Readings

- **Theory:**
  - Foster Provost and Tom Fawcett "Data Science for Business: Fundamental principles of data mining and data analytic thinking", O'Reilly Media, 2013
  - Jeffrey Stanton "Introduction to Data Science", 2012.
    https://storage2.ischool.syr.edu/media.ischool.syr.edu/oldmedia/documents/2012/3/DataScienceBook1_1.pdf
  - Lars Nielsen, Noreen Burlingame "A simple introduction to data science", 2013 (ultra-short introduction)
  - Emmanuel Ameisen "Building Machine Learning Powered Applications", O'Reilly, 2020, https://www.oreilly.com/library/view/building-machine-learning/9781492045106/
  - Rachel Schutt "Doing data science", O'Reilly 2013
  - Jiawei Han "Data Mining: Concepts and Techniques", 3rd edition 2012.
  - Kirill Dubovikov "Managing Data Science: Effective strategies to manage data science projects and build a sustainable team", Packt Publishing, 2019
  - José Hernández-Orallo, M.José Ramírez-Quintana, Cèsar Ferri, "Introducción a la minería de datos", Pearson 2004
  - Peter Flach "Machine learning: the art and science of algorithms that make sense of data", Cambridge 2013.
  - Martínez-Plumed, F., Contreras-Ochando, L., Ferri, L., Hernandez Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M.J., and Flach, P. "CRISP-DM twenty years later: From data mining processes to data science trajectories." IEEE Transactions on Knowledge and Data Engineering (2019).
- **Lab (R and Python):**
  - CRAN manuals: http://cran.r-project.org/doc/manuals/R-intro.pdf 2020, http://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf, 2014.
  - Luis Torgo "Data Mining with R", CRC Press 2010.
  - Wikibooks: http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R,2019 http://en.wikibooks.org/wiki/R_Programming, 2019.
  - Graham Williams: Hands-On Data Science with R, http://onepager.togaware.com/
  - Wes McKinney "Python for Data Analysis Data Wrangling with Pandas, NumPy, and Ipython"
  - Toby Segaran "Programming Collective Intelligence: Building Smart Web 2.0 Applications", 2007
  - Raúl Garreta, Guillermo Moncecchi "Learning scikit-learn: Machine Learning in Python" 2013