Data Science Project III

14021 – PRYIII – "Proyecto III, análisis de datos"

2024-2025

- José Hernández Orallo, DSIC, UPV, jorallo@upv.es
- Joaquín Martínez Minaya , DEIOAC, UPV, jmarmin@eio.upv.es
- J. Alberto Conejero, DMAT, UPV, aconejero@upv.es

Description

- Credits: 6.0 (3: seminar, 3: lab)
- Lecturers
  - José Hernández Orallo (jorallo@upv.es)
    - Office 236, 2nd floor DSIC (Bldg. 1F).
    - Attention/tutoring hours: on demand by email.
  - Joaquín Martínez Minaya (jmarmin@eio.upv.es)
    - Office Building 7A, fifth floor
    - Attention/tutoring hours: on demand by email.
  - J. Alberto Conejero (aconejero@upv.es)
    - Office: Building 1h, third floor
    - Attention/tutoring hours: on demand by email.

Learning Outcomes

After completion of the course, the student will be able to convert data into value through data-driven products. The focus will be put on data analysis, and in the context of work team in a new domain.

- Goals:
  1. recognise the value of data and the business opportunities for the development of data-driven products.
  2. estimate the complexity and resources that are needed for a data analysis project and establish the measures of cost and success.
  3. develop a data science project in a team on a new domain, learning and integrating new knowledge and technologies as needed.
  4. develop descriptive and predictive models and their evaluation in real situations (non-toy domains).
  5. communicate the results in written and oral form, and transmit a narrative of the data and its value.

Contents

- **Unit 1**: The context of a data analysis project: Opportunities, problems and success criteria
- **Unit 2**: Profiles of data science projects: business, social and research
- **Unit 3**: Task identification and model building
- **Unit 4**: Iterative improvement
- **Unit 5**: Exploitation and value assessment
- **Unit 6**: Presentation and report

Evaluation

- All around **the project**
  - Teams of ~**six students**.
  - Develop the idea of a new product from the use of data (open data, Internet, repositories, etc.) or derive data-driven knowledge that could improve an existing procedure.
- Components:
  - Portfolio: The portfolio that will be maintained weekly and inspected (through short presentations, interviews or questions) at different milestones
    - **M1 (15%), M2 (15%),** with a public rubric for each of them.
  - Final oral presentation with a report (consolidating the portfolio):
    - **Presentation plus accompanying report: G1 (60%)**
      - Rubric: data value, alternatives and innovation, technical tool integration, project effort and exposition quality.
    - **Intra-team co-evaluation questionnaire: C1 (0-1)** will multiply G1 as a coefficient
      - Rubric: percentage of contribution, disposition
    - **Pre-evaluation (rehearsal, compulsory) and final evaluation (resit) weeks**.
  - The inter-team evaluation
    - **Each group is evaluated by the rest of the class (C2: 10%),**
    - Questionnaire with a finite number of points to assign among the other groups, with a score that is obtained as a median

Grade (0-10) = M1 (0-1.5) + M2 (0-1.5) + G1 (0-6) * C1 (0-1) + C2 (0-1)

Evaluation: Transversal Competencies

- CT01 – Social and environmental compromise
  - Evaluated by rubrics in M1 (and some items of M2)
- CT02 – Innovation and Creativity
  - Evaluated by rubrics in M1 and M2 (value proposition, business canvas, etc.)
- CT03 – Working in teams and leadership
  - Evaluated by rubric C1

Use of AI

- Use of AI in Data Science is accelerating:

Can language models automate data wrangling?

Gonzalo Jaimovitch-López[1] · Cèsar Ferri[1] · José Hernández-Orallo[1] · Fernando Martínez-Plumed[1] · María José Ramírez-Quintana[1]

Received: 25 January 2022 / Revised: 16 August 2022 / Accepted: 29 September 2022 /
Published online: 1 December 2022

Machine Learning,
https://link.springer.com/article/10.1007/s10994-022-06259-9

Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering

Noah Hollmann
University of Freiburg
Charité Hospital Berlin
Prior Labs
noah.homa@gmail.com

Samuel Müller
University of Freiburg
Prior Labs
muellesa@cs.uni-freiburg.de

Frank Hutter
University of Freiburg
Prior Labs
fh@cs.uni-freiburg.de

Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track

15 Oct 2024

DATA INTERPRETER:
AN LLM AGENT FOR DATA SCIENCE

Sirui Hong[1*], Yizhang Lin[1*], Bang Liu[2§], Bangbang Liu[1], Binhao Wu[1], Ceyao Zhang[3], Chenxing Wei[4], Danyang Li[1], Jiaqi Chen[5], Jiayi Zhang[6], Jinlin Wang[1], Li Zhang[5], Lingyao Zhang[1], Min Yang[7], Mingchen Zhuge[8], Taicheng Guo[9], Tuo Zhou[10], Wei Tao[5], Xiangru Tang[11], Xiangtao Lu[1], Xiawu Zheng[12], Xinbing Liang[1,13], Yaying Fei[14], Yuheng Cheng[3], Zhibin Gou[15], Zongze Xu[16], Chenglin Wu[1§]

[1]DeepWisdom, [2]Université de Montréal & Mila, [3]The Chinese University of Hong Kong, Shenzhen, [4]Shenzhen University, [5]Fudan University, [6]Renmin University of China, [7]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, [8]King Abdullah University of Science and Technology (KAUST), [9]University of Notre Dame, [10]The University of Hong Kong, [11]Yale University, [12]Xiamen University, [13]East China Normal University, [14]Beijing University of Technology, [15]Tsinghua University, [16]Hohai University

https://arxiv.org/abs/2402.18679

- In this course we expect you to use AI as much as possible
  - But follow the Honesty Rules (Poliformat)

Course plan

Joaquín	3CDA1	Mon 9:00-11:00 LAB 1E 1.0 (von Neumann), Fri 11:30-13:30: LAB DSIC 9
Alberto	3CDA2	Mon 12:30-14:30 LAB 1E 1.0 (von Neumann), Fri 9:30-11:30: LAB DSIC 9
Jose	3CDB1	Mon 16:00-18:00 LAB 1E 2.1 (Boole), Fri 17-19 - LAB 1E 1.0 (von Neumann)
Jose	3CDB2	Mon 18:00-20:00 LAB 1E 1.0 (von Neumann), Fri 15-17 - LAB 1E 1.0 (von Neumann)

	Monday Seminars						Friday Practice					
	Sem-A1	Sem-A2	Sem-B1	Sem-B2			Lab-A2	Lab-A1	Lab-B2	Lab-B1		During the week...
	9:00-11:00	12:30-14:30	16:00-18:00	18:00-20:00	During the class (instructors-led in black, students-led in blue)		9:30-11:30	12:00-14:00	15:00-17:00	17:00-19:00	During the class we do:	Students DO or DELIVER
Feb-03	Jose	Jose	Jose	Jose	Course intro (45 min) - Look at project list - Unit1a - Value of data. The Data Sci	Feb-07	Not scheduled - No Class					Students fill and deliver form (by Thursday)
Feb-10	Jose	Jose	Jose	Jose	Unit1b - Costs (20 min) - Unit2a - DS Teams (20 min) - First group meeting (60	Feb-14	Alberto	Joaquín	Jose	Jose	Separate team meetings - instructor rounds	First group meetings with tutor - explore project
Feb-17	Alberto	Alberto	Alberto	Alberto	Unit2 - Society-oriented DS - Unit6 - Presentation and Report - Separate team m	Feb-21	Alberto	Joaquín	Jose	Jose	Separate team meetings - instructor rounds	Explore data/goals
Feb-24	Joaquín	Joaquín	Joaquín	Joaquín	Unit2 - Research-oriented DS - Rubric for M1. Separate team meetings and rour	Feb-28	Alberto	Joaquín	Jose	Jose	Separate team meetings - instructor rounds	Prepare M1
Mar-03	Jose	Jose	Jose	Jose	Unit2 - Business-oriented DS - Joint session. Separate team meetings and round	Mar-07	Alberto	Joaquín	Jose	Jose	Separate team meetings - instructor rounds	Prepare M1
Mar-10	Joaquín	Alberto	Jose	Jose	**Mid-term Presentations** (15min + 5 min questions) + Rubric M2	Mar-14	Alberto	Joaquín	Joaquín	Joaquín	Separate team meetings - instructor rounds	**MILESTONE M1 (10 March): PLAN: Team roles,**
Mar-17	Falles	Falles	Falles	Falles	Falles	Mar-21	Alberto	Joaquín	Jose	Jose	Separate team meetings - instructor rounds	Prepare M2
Mar-24	Jose	Jose	Jose	Jose	Unit3 - Task Identification - model building - Guidelines for M2's delivery - Sepa	Mar-28	Joaquín	Joaquín	Jose	Jose	Separate team meetings - instructor rounds	Prepare M2 - models
Mar-31	Joaquín	Joaquín	Joaquín	Joaquín	Unit4 - Iterative improvement - Separate team meetings and rounds	Apr-04	Exam Perio	Exam Period	Exam Period	Exam Period	Exam Period	
Apr-07	Exam Perio	Exam Period	Exam Period	Exam Period	Exam Period	Apr-11	Alberto	Joaquín	Jose	Jose	Separate team meetings - instructor rounds	Prepare M2
Apr-14	Jose	Jose	Jose	Jose	Unit5 - Exploitation and value assessment - Separate team meetings and rounds	Apr-16	Alberto	Joaquín	Jose	Jose	Separate team meetings - instructor rounds	**MILESTONE M2 (16 April): MOCKUP: minable v**
Apr-21	Easter	Easter	Easter	Easter	Easter	Apr-25	Easter	Easter	Easter	Easter	Easter	
Apr-28	Easter	Easter	Easter	Easter	Easter	May-02	Alberto	Joaquín	Jose	Jose	Separate team meetings - instructor rounds	Iterate on models. Prepare final presentation
May-05	Joaquín	Joaquín	Joaquín	Joaquín	Unit6 - Presentation and report - rubric for G1 - Separate team meetings and ro	May-09	Ada Byron	Ada Byron's	Ada Byron's	Ada Byron's day		
May-12	Joaquín	Alberto	Jose	Jose	**FINAL PRESENTATION: G1 Rehearsals. Students present and fill co-e**	May-16	Alberto	Joaquín	Jose	Jose	**FINAL PRESENTATION: G1 Rehearsals. Stud**	G1: Report Delivery - Improvements and final details
May-19	Joaquín	Alberto	Jose	Jose	Separate team meetings - instructor rounds	May-23	Alberto	Joaquín	Jose	Jose	Separate team meetings - instructor rounds	
May-26	Joaquín	Alberto	Jose	Jose	**FINAL PRESENTATION: G1 Final**	May-30	Alberto	Joaquín	Jose	Jose	**FINAL PRESENTATION: G1 Final**	**G1: Report resit - FAIR**
						May-29 (Thursday) is the Project Fair						

Will probably change during the term, so please refresh to new versions on poliformat frequently.

Project Fair

https://www.feriaetsinf.org/

Tentative: final info will be given when details are fully clear

- **May-29 (Thursday):**
  - **9:00 registration, Year 2: 9:00-11:00, Year 3: 12:00-14:00 (posters to be delivered about 10 days before).**
    - A minimum number of students per team have to go to the fair (from 9 o 14) and be at the stand from 12:00h to 14:00h
    - Teams do a **project pitch** in ~15min using a computer (demo, slides, etc.) to companies coming around.
- Effect on the grades
  - 1) If the team **does the retake** to improve their grades from the previous week, they will need to **tell us in advance**, so the tutor (Jose, Joaquín or Alberto) will be there to listen the project pitch. Any pitch to a company can be reused for this, in order to avoid many repetitions.
  - 2) Independently of **doing or not the retake**: Being at the fair and placing a good poster could give **up to an extra point (+1)** for the project grade (G1).

Recommended Readings

General:

- Kirill Dubovikov "Managing Data Science: Effective strategies to manage data science projects and build a sustainable team", Packt Publishing, 2019.
- Provost and Fawcett "Data science for business: what you need to know about data mining and data-analytic thinking", O'Reilly, 2013.
- Carl Anderson "Creating a Data-Driven Organization: Practical Advice from the Trenches" O'Reilly, 2015.
- de Graaf "Managing Your Data Science Projects: Learn Salesmanship, Presentation, and Maintenance of Completed Models", Apress 2019.
- Schutt and O'Neil "Doing data science: Straight Talk from the Frontline" O'Reilly 2013
- Emmanuel Ameisen "Building Machine Learning Powered Applications", O'Reilly, 2020, https://www.oreilly.com/library/view/building-machine-learning/9781492045106/ (chapters 1-6, useful for Units 1, 3, 4, 5)
- Efron, Bradley, and Trevor Hastie. *Computer age statistical inference*. Vol. 5. Cambridge University Press, 2016.
- Osterwalder, Alexander, et al. *Value proposition design: How to create products and services customers want*. John Wiley & Sons, 2014. (Unit 2 – business)

We will provide more specific pointers for each unit.