Data Science (CDA)

Project Guidelines

- 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

Project Guidelines

- Freelance data scientist project: G1 (50%) * C1 (0..1)
  - Groups of 3 students with co-evaluation.
  - 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: innovation, tool integration, data value, presentation.
  - Co-evaluation rubric: percentage of work (not hours) of your peers, collaborative attitude, resolution, disposition.

Project Guidelines

- Imagine a problem or start with some data.
  - Examples:
    - https://www.projectpro.io/projects/data-science-projects
    - https://medium.com/coders-camp/180-data-science-and-machine-learning-projects-with-python-6191bc7b9db9
    - https://towardsdatascience.com/8-data-science-project-ideas-from-kaggle-in-2021-83a3660e0342
    - Poliformat

Project Guidelines

- Imagine a problem or start with some data.
  - Data sources (it can be in->in, in->out, out->in, out->out or ∅→out):
    - Your own data (personal or business)
    - Your own interests (games, music, …)
    - https://datasetsearch.research.google.com/
    - Social media
    - http://blog.bigml.com/2013/02/28/data-data-data-thousands-of-public-data-sources/
    - http://www.valencia.es/ayuntamiento/datosabiertos.nsf/fCategoriaVistaAcc_busqueda?ReadForm&lang=1&nivel=2&Vista=vCategoriasAccTodas&Categoria=Sin_categoria&idapoyo=22ADF97C1FD223B5C1257C55003BD01F
    - http://datosabiertos.malaga.eu/
    - https://www.reddit.com/r/datasets/
    - https://www.kaggle.com
    - http://archive.ics.uci.edu/ml/
    - https://www.citibikenyc.com/system-data
    - http://users.dsic.upv.es/~flip/BikeSharingDemand/

Project Guidelines

- Data-driven products
  - Do you see a problem or goal that can be solved or improved with the data?
    - Is it novel?
    - Would the solution lead to added value?
    - Would it possibly lead to a product?
    - Would it lead to economic return?

Project Guidelines

- Once you have a domain, some data and a data-driven goal, then you can analyse the data!
  - Apply visualisation and modelling techniques to the data in order to see whether the goals can be achieved.
  - You can use R, Python or any other tool.
  - Explore, report, play, …
  - Be bold, be insightful!

Project Guidelines

- Guidelines for the oral presentation
  - You can use powerpoint or any other presentation tool
  - Do you transmit the ideas easily?
  - Is visualisation used so that people can understand the data and the goals?
  - Tell a story:
    - http://www.techrepublic.com/article/be-the-hemingway-of-data-science-storytelling/
    - https://hbr.org/2015/10/the-best-data-scientists-know-how-to-tell-stories

Project Guidelines

- Recommended timing
  - Early October: create a group and have a first meeting.
  - Mid October: decide the domain or data to work with.
  - Late October: download data, clarify goals, possible products
  - Early November: visualise and analyse data
  - Late November: prepare a first version of the presentation
  - (Nov-28 – Nov-30) Week for rehearsing the presentations (PreEvaluation): make your first group presentation.
  - Polish presentation, make a convincing story.
  - (Dec-19 – Dec-21) Second week for Final Evaluation.

Project Guidelines

- RUBRIC FOR EVALUATION (0-5)
  - **Data VALUE** (0-1): The students have identified the value of the data they have worked with, its applicability in their contemporary world, the people who may be benefited by this work, and possible apps or even a future entrepreneurship as an outcome. They have also identified the limitations and sustainability issues, as well as the overall impact of the use of the data and the proposed idea on society.
  - **ALTERNATIVES and innovation** (0-1): The students have looked for alternative proposals (bibliography, websites, apps) for the same domain, the same data or application, and have compared their results with them (at least at the abstract level), and see whether what they present is the same or innovative, is below the current state of the art, covers real needs, etc. Preliminary market studies are well received.
  - **TECHNICAL tool integration** (0-1): The students have mastered different technical tools and their integration and have used appropriately for their goals. The principles and expertise seen during the course are reflected in the technical solutions.
  - **Project EFFORT** (0-1): The amount of work that the project required, including data collection and integration effort, and the lessons learnt and how they have solved the problems or found a way around. Evidence of teamwork.
  - **EXPOSITION quality** (0-1): The quality of the presentation, the students have been able to transmit an idea, clearly, the motivation of the work and the insights. The quality of the slides and the graphics are impeccable. They make gestures, avoid being monotonous. They are telling a story. They answer the questions precisely.

Be original!

Project Guidelines

- Instructions for the presentation:
  - This year presentations will have 15 minutes (strict) + 5 minutes for questions.
  - Each group will book a day for the rehearsal and, if so decided, the final presentation, in the corresponding weeks, by email.
  - Before the presentation, the members will send an email with a few lines describing what part of the work has been done by each member of the team (or percentages).
  - After the presentation, the students will send the slides (no other report or material is needed).

Project Guidelines

- Important: Note about teamwork:
  - If there are disagreements inside a group, try to resolve them internally. This is part of what makes working in groups important.
  - If during the development of the project there are big issues (e.g., one member doesn't work at all, agreements are impossible, etc.) then please let us know. In extreme cases, we will split the group.
  - Don't use other partners as excuses (e.g., this doesn't work because "X didn't do his part"). Even in difficult situations and disagreements, you have to give priority to the common goal.

Project Guidelines

- RUBRIC FOR CO-EVALUATION (0-1)
  - Evaluate each teammate X including yourself:
    - **C: Contribution (0-100):** What's the percentage of the total contribution that can be attributed to X (consider the result of the project, not hours of work, as some people are more efficient than others).
    - **D: Disposition (0-100):** On a scale between 0 and 100, how would you value the collaborative attitude of your teammate X (disposition, helpfulness, seeking consensus rather than conflicts, etc.).
  - The sum of "contribution" for all X in the team should be 100.
- Given the C and D values for all teammates in a group, we will use a procedure (not disclosed to avoid optimising for it instead of being honest), to derive a score between 0 and 1 for each teammate, according to the values and harmony of these assessments.