Universitat Autònoma de Barcelona

**2022/2023**

# Data Engineering

Code: 106565
ECTS Credits: 6

| Degree | Type | Year | Semester |
|---|---|---|---|
| 2504392 Artificial Intelligence | OB | 1 | 2 |

## Contact

Name: Javier Vazquez Corral
Email: javier.vazquez.corral@uab.cat

## Use of Languages

Principal working language: english (eng)
Some groups entirely in English: Yes
Some groups entirely in Catalan: No
Some groups entirely in Spanish: No

## Prerequisites

Students should have cursed and understood the subject of "Introduction to Programming I" and "Mathematical Foundations I".

## Objectives and Contextualisation

The subject aims to provide the fundamentals of data analysis and visualization. The different stages of data analysis processes will be studied, from the collection, annotation and preparation of data, to its analysis and visualization, preparing the way for a more advanced modelling through Machine Learning (M8).

## Competences

- Conceptualize and model alternatives of complex solutions to problems of application of artificial intelligence in different fields and create prototypes that demonstrate the validity of the proposed system.
- Introduce changes to methods and processes in the field of knowledge in order to provide innovative responses to society's needs and demands.
- Know and efficiently use techniques and tools for representation, manipulation, analysis and management of large-scale data.
- Know, understand, use and apply appropriately the mathematical foundations necessary to develop systems for reasoning, learning and data manipulation.
- Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.
- Work cooperatively to achieve common objectives, assuming own responsibility and respecting the role of the different members of the team.

## Learning Outcomes

1. Conceive, design and implement data collection and annotation processes appropriate to the problem at hand that needs resolving.
2. Properly use data visualization methods.
3. Propose new methods or informed alternative solutions.

4. Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.
5. Understand how to use the basic tools for manipulating different types of structured, semi-structured and unstructured data.
6. Use data analysis methods to test hypotheses and obtain useful interpretations.
7. Work cooperatively to achieve common objectives, assuming own responsibility and respecting the role of the different members of the team.

## Content

This subject will be divided into 11 main topics:

- Introduction.

- Basic statistics Recap.

- Numpy / Matplotlib/ Pandas
-XML/JSON

- Regular Expressions/DFA/NFA

- Data types. Missing data.

- Introduction to data analysis. Outliers.

- PCA

- kNN

- Image retrieval and recommender systems.

- K-means

## Methodology

There will be three types of sessions:

Theory classes: The objective of these sessions is for the teacher to explain the theoretical background of the subject. For each one of the topics studied, the theory and mathematical formulation is explained, as well as the corresponding algorithmic solutions.

Exercise sessions: They will be sessions that facilitate interaction. In these sessions, the aim is to reinforce the comprehension of the topics seen in the theory classes by proposing practical cases that require the design of a solution in which the methods seen in the theory classes are used.

Practical laboratory sessions: They will be sessions in which different types of activities related to the realization of the project/projects by teams of students will be carried out. During the practical sessions the project/projects to be solved will be presented and a series of activities will be carried out in teams of students in collaborative work mode. The identification of the problem, the discussion of the design, the distribution and organization of the work to be carried out, the development of the solution and the presentation of the results to the teacher and the rest of the students will be addressed.

All the information of the subject and the related documents that the students need will be found in the virtual campus.

Within the schedule set by the centre or degree programme, 15 minutes of one class will be reserved for students to evaluate their lecturers and their courses or modules through questionnaires.

Annotation: Within the schedule set by the centre or degree programme, 15 minutes of one class will be reserved for students to evaluate their lecturers and their courses or modules through questionnaires.

## Activities

| Title | Hours | ECTS | Learning Outcomes |
|---|---|---|---|
| Type: Directed | | | |
| Exercise classes | 12 | 0.48 | 5, 2, 6 |
| Theory classes | 28 | 1.12 | 5, 3, 6 |
| Type: Supervised | | | |
| Lab classres | 12 | 0.48 | 1, 5, 3, 4, 7, 2, 6 |
| Type: Autonomous | | | |
| Completion of the exercises at home | 16 | 0.64 | |
| Project labs | 58 | 2.32 | 1, 5, 3, 4, 7, 2, 6 |
| Study of the theory part | 20 | 0.8 | 5 |

## Assessment

Final mark:

The final grade is calculated using a weighted average according to the different activities that are carried out:

Final grade = 0.4 * Theory Grade + 0.2 * Exercises Grade + 0.4 * Laboratory Grade

For applying this formula, the condition is that both the theory and the laboratory grades, are higher than 5. When a student does not reach the minimum required in some one of the evaluation activities, then the final grade will be that of the element not allowing the calculation (i.e. if a student has a 6 in Theory, a 5 in Exercises, but a 2 in Laboratory, the final mark will be a 2).

Theory Grade

The theory grade aims to assess the individual abilities of the student in terms of the theoretical content of the subject. This is done continuously during the course through two partial exams:

Theory Grade = 0.5 * Grade Exam 1 + 0.5 * Grade Exam 2

Exam 1 is done in the middle of the semester and serves to eliminate part of the subject if it is passed. Exam 2 is done at the end of the semester and serves to eliminate part of the subject if it is passed. These exams aim to assess the abilities of each student in an individualized manner, both in terms of solving exercises using the techniques explained in class, as well as evaluating the level of conceptualization that the student has made of the techniques seen. In order to obtain a final pass theory grade, it will be required for the partial exam grades 1 and 2 to be both higher than 4.

Recovery exam. In case the theory grade does not reach the adequate level to pass, the students can take a recovery exam, destined to recover the failed part (1, 2 or both) of the continuous evaluation process. Note: The recovery exam can also be taken if the student wants to obtain a higher grade than in the previous ones.

Exercises Grade

The aim of the exercises is for the student to train with the contents of the subject continuously and become familiar with the application of the theoretical concepts. As evidence of this work, the presentation of a portfolio is requested in which the exercises worked out will be collated:

Exercises Grade = Portfolio evaluation

Practical Laboratory Sessions Grade

The part of laboratory based practical sessions carries an essential weight in the overall mark of the subject. Laboratory sessions aim for the student to design a solution to a problem that is set out in a contextualized way. Such problems will require the design of an integral solution, from the exploration of available techniques to data modelling. In addition, the students must demonstrate their teamwork skills and present the results to the class convincingly. Laboratory sessions are structured around project/s. The projects are evaluated through its deliverable, an oral presentation that students will make in class, and a self-evaluation process. The grade is calculated as follows:

Project Grade = 0.6 * Grade Deliverables + 0.3 * Grade Presentation + 0.1 * Grade Self-evaluation

In case of being more than 1 project, the Laboratory Grade will be the average of all the Different Project Grades. In case of not passing the project, the recovery of the part of the deliverables of the unsuccessful projects will be allowed, restricted to a maximum grade of 7/10. The oral presentation cannot be recovered.

## Assessment Activities

| Title | Weighting | Hours | ECTS | Learning Outcomes |
|---|---|---|---|---|
| Partial exam 1 | 0.2 | 2 | 0.08 | 1, 5, 3, 4, 6 |
| Partial exam 2 | 0.2 | 2 | 0.08 | 1, 5, 3, 4, 6 |
| Portfolio | 0.2 | 0 | 0 | 1, 5, 3, 4, 2, 6 |
| Practical project | 0.4 | 0 | 0 | 1, 5, 3, 4, 7, 2, 6 |

## Bibliography

Data Science from Scratch: First Principles with Python, Joel Grus, O'Reilly Media, 2015, 1st Ed.

Python Data Science Handbook, Jake Van der Plas, O'Reilly Media, 2016, 1st Ed.

Computational and Inferential Thinking: The Foundations of Data Science, Ani Adhikari and John DeNero, online: https://ds8.gitbooks.io/textbook/content/

## Software

This subject will use Python as the programming Language.