COURSE DATA

DATA SUBJECT

Code: 36423
Name: Data processing
Cycle: Undergraduate Studies
ECTS Credits: 6
Academic year: 2026-27

STUDY (S)

Degree	Center	Acad. year	Period
1406 - Degree in Data Science	Escola Tècnica Superior d'Enginyeria	1	Second quarter

SUBJECT-MATTER

Degree	Subject-matter	Character
1406 - Degree in Data Science	Information Management	COMPULSORY

COORDINATION

MARTINEZ SOBER MARCELINO

GOMEZ CHOVA LUIS

SUMMARY

The data scientist is faced with a dataset of very different origin, format, organisation, coding, etc. The correct acquisition, organisation, elimination of possible erroneous data (outliers), missing data imputation, data transformation, selection of the most relevant characteristics of a high dimensionality set (feature selection), elimination of redundant data, etc. is one of the most costly stages of a data analysis problem. This stage is crucial for the correct treatment of the problem and the reliability and solidity of the results obtained in later stages of analysis (selection of models, classifiers, grouping, estimation, hypothesis contrasts, etc.). All these tasks will be dealt with in the compulsory subject 36423 Data Processing, which is taught in the second term of the first year.

Theory lessons will be taught in Spanish and practical and laboratory lessons as according to the information sheet available on the web page of the degree.

PREVIOUS KNOWLEDGE

RELATIONSHIP TO OTHER SUBJECTS OF THE SAME DEGREE

There are no specified enrollment restrictions with other subjects of the curriculum.

OTHER REQUIREMENTS

It is recommended to have passed the subject Data, Science and Society taught in the first term of the first course of the degree.

COMPETENCES / LEARNING OUTCOMES

1406 - Degree in Data Science

(CB3) Students must have the ability to gather and interpret relevant data (usually in their field of study) to make judgements that take relevant social, scientific or ethical issues into consideration.

(CE02) To methodologically know and apply the programming techniques and the algorithms necessary for the efficient processing of information and the computer resolution of problems that use large volumes of data.

(CE06) Ability to represent and visualise data sets for the extraction of knowledge.

(CE11) Ability to design and implement data acquisition, its integration, transformation, selection, verification of its quality and veracity from different sources, taking into account its character, heterogeneity and variability.

(CE13) To know how to design, apply and evaluate data science algorithms for the resolution of complex problems.

(CG06) Ability to access and manage information in different formats for subsequent analysis in order to obtain knowledge from data.

(CT03) Ability to defend your own work with rigor and arguments and to expose it in an adequate and accurate way with the use of the necessary means.

(CT04) To be responsible for ones own professional development and specialisation, applying the acquired knowledge in the identification of career opportunities and sources of employment.

DESCRIPTION OF CONTENTS

1. Introduction to data processing

1.1. Why analyse data?
1.2 . Overview of a data processing problem.

2. Getting data

2.1. Introduction.
2.2. Getting data.
2.3. Data repositories.
2.4 Data file formats.
2.5. Merging data from different sources.
2.6. Access to databases.

3. Data visualisation

3.1. Explanatory and exploratory graphics.
3.2. Graphic systems in R: base, grid, lattice, ggplot2.
3.3. ggplot2 library: basic representations.

4. Preparation of the data

4.1 Structure of a data set for analysis: basic data operations.
4.2. Data manipulation. tidyr library.
4.3. Data handling. dplyr library.

5. Exploratory data analysis I. Definitions

5.1. Exploring a new dataset.
5.2. Characterisation of variables.
5.3. Displaying relationships between variables.

6. Exploratory Data Analysis II. Abnormalities.

6.1. Anomalies in numerical variables: Outliers. Characterisation of outliers. Detection methods.
6.2. Abnormalities in numerical variables: missing and absent data.

7. Working with text data

7.1. Introduction
7.2. Bases of text data analysis.
7.3. Basic functions for handling characters in R.
7.4. Regular expressions.
7.5 Character encoding: ascii vs. Unicode.

8. Data processing practices

In this block a series of practical cases will be presented as laboratory practices.
Practice 1. Data import.
Practice 2. Data visualization.
Practice 3. Preparation of data with tidyr.
Practice 4. Data management with dplyr.
Practice 5. Anomaly detection in data.
Practice 6. Exploratory data analysis.
Practice 7. Complete analysis of a data set.

WORKLOAD

PRESENCIAL ACTIVITIES

Activity	Hours
Theory	34,00
Laboratory	20,00
Classroom practices	6,00
Total hours	**60,00**

NON PRESENCIAL ACTIVITIES

Activity	Hours
Attendance at other activities	0,00
Individual or group project	15,00
Independent study and work	20,00
Preparation of lessons	30,00
Preparation for assessment activities	25,00
Resolution of case studies	0,00
Total hours	**90,00**

TEACHING METHODOLOGY

Lessons will combine theoretical and practical content.

MD1 - **Theoretical activities**. Expositive development of the subject with the participation of the student in the resolution of specific issues. Individual evaluation questionnaires.

In the in class theoretical activities the topics of the subject will be developed providing a global and integrating vision, analysing in greater detail the key aspects and of greater complexity, encouraging, at all times, the participation of the students (CB03, CT03).

MD2 - **Practical activities**. Learning by solving problems, exercises and case studies through which competences are acquired on the different aspects of the subject. (CB03, CG06, CE02, CE06, CE11, CE13)

The theoretical activities are complemented by practical activities with the aim of applying the basic concepts and expanding them with the knowledge and experience that are acquired during the realisation of the proposed works.

MD4 -**Work in the laboratory and / or computer classroom**. Learning by carrying out activities developed individually or in small groups and carried out in laboratories and / or computer classrooms. (CB03, CG06, CE02, CE06, CE11, CE13)

In addition to in class activities, students must perform personal tasks (outside the classroom) on issues and problems, as well as the preparation of classes and exams (study). These tasks will be carried out mainly individually, in order to enhance autonomous work, but will also include work, especially the preparation and resolution of laboratory practices, which require the participation of small groups of students (2-3) to promote ability to integrate into work groups.

The e-learning platforms (Virtual Classroom) of the Universitat de València, Microsoft Teams and Blackboard Collaborate, will be used as a communication support with students. Through the students will have access to the didactic material used in class, as well as the problems and exercises to solve.

EVALUATION

The learning of the knowledge and competences achieved by students will be continuously evaluated throughout the course, and will consist of the following evaluation blocks:

First and second exam calls

SE1 - Objective test, consisting of one or more examinations with both theoretical and practical questions and problems (53%) **(Note: All percentages refer to the final grade)** (evaluation of competencies CB03, CG06, CT03, CE02, CE06, CE11, CE13)

- SE1-1 (35%) Theory examination

- SE1-2 (18%) Laboratory examination

SE2 - Evaluation of practical activities based on the preparation of papers/memories and/or oral presentations (evaluation of competences CB03, CG06, CT03, CT04, CE02, CE06, CE11, CE13) (27%)

- SE2-1 (15%) Implementation of a mini-project consisting of an introduction to the Data Science stages.

- SE2-2 (12%) Laboratory lessons. (NON-recoverable activity)

SE3 - Continuous assessment of each student, based on the participation and degree of involvement of the student in the teaching-learning process, taking into account the regular attendance to the planned face-to-

face activities and the resolution of issues and problems proposed periodically. (20%)

- SE3-1 (5%) Regular attendance at planned face-to-face activities. (NON-recoverable activity) (evaluation of competencies CB04, CG01)

- SE3-2 (15%) Resolution of proposed issues and problems. (NON-recoverable activity) (evaluation of competencies CB02, CB04, CG01, CT03)

The final grade of the course will be calculated as the weighted average of each of the previous sections, according to the following criteria: SE-1 (53%), SE-2 (27%), SE-3 (20%).

Particular considerations about the evaluation:

- A minimum score of 5 (out of 10) is required in the evaluation sections SE1-1, SE1-2+SE2-2 and SE2-1.

- The SE2-2, SE3-1 and SE3-2 activities are not recoverable.

- Activity SE1-2 will take place at the end of the theory exam on the day of the official call.

In any case, the evaluation system will be governed by the provisions of the Evaluation and Qualification Regulations of the University of Valencia for Degrees and Masters:

https://webges.uv.es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=start&idEdictSelected=5639

Copying or plagiarism of any activity that is part of the evaluation will result in the impossibility of passing the course, and the student will then be subject to the appropriate disciplinary procedures indicated in the ACTION PROTOCOL FOR FRAUDULENT PRACTICES AT THE UNIVERSITY OF VALENCIA (ACGUV 123/2020).

REFERENCES

- R.K.Pearson (2018) Exploratory Data Analysis Using R. CRC.
- H. Wickham, Mine Çetinkaya-Rundel, G. Grolemund. (2023) R for data Science 2nd Edition. O¿Reilly Media Inc. http://r4ds.had.nz/
- B. S. Baumer, D. T. Kaplan, N. J. Horton (2021) Modern Data Science with R. 2nd Edition. Boca Raton : Taylor & Francis CRC Press. (disponible e-libro)
- R. Buttres y, L.R. Whitaker (2018). A data scientist's guide to acquiring, cleaning and managing data in R . Wiley. (disponible e-libro)
- W. Graham, (2017). The Essentials of Data Science: Knowledge Discovery Using R. Chapman and Hall/CRC. (disponible e-libro)
- L. Han, M. Kamber, and J. Pei. (2012) Data Mining Concepts and Techniques (third Edition).

(disponible e-libro)
- N. Zumel and J. Mount (2014). Practical Data Science with R. Manning Publications Co.
- A.Cirillo (2017) R Data Mining. Pack Publishing (disponible e-libro)
- C.Aggarwal (2015) Data mining: the textbook. Springer (disponible e-libro)