COURSE DATA

DATA SUBJECT

Code: 46573
Name: Exploratory data analysis
Cycle: Master's Degree
ECTS Credits: 4.5
Academic year: 2025-26

STUDY (S)

Degree	Center	Acad. year	Period
2262 - Master's Degree in Data Science	Escola Tècnica Superior d'Enginyeria	1	First quarter

SUBJECT-MATTER

Degree	Subject-matter	Character
2262 - Master's Degree in Data Science	Exploratory data analysis	COMPULSORY

COORDINATION

MARTINEZ SOBER MARCELINO

GOMEZ SANCHIS JUAN

SUMMARY

In this course, we describe the initial stages of a data analysis problem.

The data scientist is faced with datasets coming from very diverse sources, formats, structures, codifications, etc
acquisition, organization, removal of possible erroneous data (outliers), imputation of missing values (m
imputation), data transformation, dimensionality reduction (feature selection of the most relevant variables in high-
datasets and feature extraction), elimination of redundant data, etc., constitutes one of the most costly stage
analysis process. This stage is crucial for the proper treatment of the problem and for ensuring the reliability and ro
the results obtained in later stages of analysis (model selection, classification, clustering, estimation, hypothesis t
In this module, we will focus on the data preparation stages.

PREVIOUS KNOWLEDGE

RELATIONSHIP TO OTHER SUBJECTS OF THE SAME DEGREE

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

OTHER REQUIREMENTS

Introduction to Data Science

COMPETENCES / LEARNING OUTCOMES

2262 - Master's Degree in Data Science

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

Be able to assess the need to complete their technical, scientific, language, computer, literary, ethical, social and human education, and to organise their own learning with a high degree of autonomy.

Capacidad de análisis y síntesis, en la elaboración de informes, en la exposición, comunicación y defensa de ideas.

Capacidad de organización y planificación de actividades de investigación, desarrollo y consultoría en el área de ciencia de datos.

Entender la utilidad de la ciencia de datos y sus elementos asociados, así como su aplicación en la resolución de problemas, eligiendo las técnicas más adecuadas a cada problema, aplicando de forma correcta las técnicas de evaluación y, finalmente,
interpretando los modelos y resultados.

Extraer conocimiento de conjuntos de datos en diferentes formatos.

Ser capaces de acceder a herramientas de información (bibliográficas y de empleo) y utilizarlas apropiadamente.

Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio, aplicando los conocimientos adquiridos en la identificación de salidas profesionales y yacimientos de empleo.

Students should demonstrate self-directed learning skills for continued academic growth.

DESCRIPTION OF CONTENTS

1. Introduction to exploratory data analysis

In this block an introduction showing the main aspects of data visualization will be done in order to get a correct data visualization.

2. Getting and cleaning data

In this block the different data types (continuous , discrete) , importing data stored in the most common formats , data conversion , detection of anomalous data will be presented.

3. Statistical data analysis

In this block, a first approach to statistical and visual data analysis is presented. This task is a fundamental part in the understanding of the available data and in the detection of wrong values (univariate , bivariate and multivariate analysis , correlation , covariance , etc.)

4. Exploratory Data Analysis II. Abnormalities.

This block presents different types of anomalies present in numerical data such as outliers, detection methods, as well as the problem of missing values, their types and imputation methods.

5. Data transformations

This block presents methods of data transformation. In this processing step, the data are transformed or consolidate so that the resulting mining process may be more efficient, and the patters found may be easier to understand.

6. Introduction to dimensionality reduction

In this block an introduction to dimensionality reduction techniques is presented. In particular, feature selection and feature extraction paradigms will be presented. In particular, simple filter feature selection methods and more complex wrapper methods will be presented. Similarly, the feature extraction paradigm, in particular principal component analysis (PCA), will be presented.

WORKLOAD

PRESENCIAL ACTIVITIES

Activity	Hours
Theory	25,00
Theoretical and practical classes	4,00
Laboratory	16,00
Total hours	**45,00**

NON PRESENCIAL ACTIVITIES

Activity	Hours
Attendance at other activities	0,00
Individual or group project	10,00
Independent study and work	7,50
Preparation of lessons	16,50
Preparation for assessment activities	6,00
Resolution of case studies	5,00
Total hours	**45,00**

TEACHING METHODOLOGY

The course will combine the theoretical and the practical part, without separating sessions devoted to theory from those devoted to practice. The lessons will be taught in a computer equipped classroom.

In the theoretical part of the classes, the teacher will introduce the concepts and methods with examples and exercises to be solved by the students.

The practical sessions will be synchronized with the theory. In these sessions, the students will learn by solving problems, exercises and case studies, in order to acquire the skills of this course.

EVALUATION

The educational evaluation of knowledge and skills achieved by the students will be made continuously throughout the course, and will consist in the following blocks of evaluation:

1. Exercises and class works submitted during the course and/or partial exams: 60% of the final grade.

2. Final exam: 40% of the final grade.

Grades obtained in paragraph 1 shall only be kept in the two examination sittings of the academic year in which they were made, since their evaluation is only possible in the teaching period

REFERENCES

- K.Pearson (2018) Exploratory Data Analysis Using R. CRC.

- H. Wickham, G. Grolemund. (2016) R for data Science. OReilly Media Inc. http://r4ds.had.co.nz/
- Max Kuhn, Kjell Johnson (2021) Feature Engineering and Selection A Practical Approach for Predictive Models-CRC Press
- GB. S. Baumer, D. T. Kaplan, N. J. Horton (2017) Modern Data Science with R. 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)
- R. D. Peng (2016) Exploratory Data Analysis with R. Lean Publishing (https://leanpub.com/exdata)
- Max Kuhn, Julia Silge (2022), Tidy Modeling with R, O'Reilly Media, Inc. https://www.tmwr.org/dimensionality.html
- Alice Zheng, Amanda Casari (2018)- Feature Engineering for Machine Learning_ Principles and Techniques for Data Scientists-OReilly Media