VNIVERSITAT Ġˀ ID VALÈNCIA

COURSE DATA

Data Subject	
Code	46573
Name	Exploratory data analysis
Cycle	Master's degree
ECTS Credits	4.5
Academic year	2023 - 2024

Study (s)			
Degree	**Center**	**Acad. year**	**Period**
2262 - Master's Degree in Data Science	School of Engineering	1	First term

Subject-matter		
Degree	**Subject-matter**	**Character**
2262 - Master's Degree in Data Science	5 - Exploratory data analysis	Obligatory

Coordination	
Name	**Department**
GOMEZ SANCHIS, JUAN	242 - Electronic Engineering
MARTINEZ SOBER, MARCELINO	242 - Electronic Engineering

SUMMARY

This course includes the first stages associated to a data analysis problem as well as the first linear statistical models related to regression and classification methods.

Data scientists usually face a set of data from different sources, format, organization, code, etc. The correct data acquisition, organization, outlier treatment, missing data imputation, data transformation, dimensionality reduction (feature selection, removal of redundant information and feature extraction) is one of the most relevant and difficult stages of data analysis. This is a key stage to perform an appropriate processing of the data and ensure reliable and robust results (model selection, classifiers, assembling, estimation, hypothesis tests, prediction, etc). This block also focuses on the subsequent stages relatives to data preparation and the statistical analysis of the data.

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 (RD 1393/2007) // LEARNING OUTCOMES (RD 822/2021)

2262 - Master's Degree in Data Science

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

- 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.

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

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

- 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.

- Extraer conocimiento de conjuntos de datos en diferentes formatos.

- 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.

LEARNING OUTCOMES (RD 1393/2007) // NO CONTENT (RD 822/2021)

- To know the techniques and algorithms in order to pre-process select and extract the more important characteristics of a set of data.
- To select the more appropriate transformations for the problem to be solved.
- To know the techniques to dataframe manipulation.
- To know the basic principles of the statistical learning to deal with missing data.
- To know the basic principles of outliers detection.
- To know the basic principles of visualization techniques in order to make exploratory data analysis.

VNIVERSITAT ID VALÈNCIA

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.

VNIVERSITAT Ð ÐVALÈNCIA

WORKLOAD

ACTIVITY	Hours	% To be attended
Theory classes	25,00	100
Laboratory practices	16,00	100
Theoretical and practical classes	4,00	100
Development of individual work	10,00	0
Study and independent work	6,00	0
Readings supplementary material	1,50	0
Preparation of evaluation activities	6,00	0
Preparing lectures	10,00	0
Preparation of practical classes and problem	6,50	0
Resolution of case studies	5,00	0
TOTAL	**90,00**	

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

Basic

- 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)

Additional

- 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