Course guide
480093 - TDS - Socio-Environmental Data Science

Last modified: 12/06/2023

Unit in charge: Barcelona School of Civil Engineering
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.

Degree: MASTER'S DEGREE IN SUSTAINABILITY SCIENCE AND TECHNOLOGY (Syllabus 2013). (Optional subject).

Academic year: 2023 **ECTS Credits:** 5.0 **Languages:** English

LECTURER

Coordinating lecturer: KARINA GIBERT OLIVERAS

Others: Karina Gibert Oliveras
Miquel Sànchez-Marrè

PRIOR SKILLS

Basics knowledge of R package
Basic programming skills
Basic Statistics

REQUIREMENTS

Fonaments d'Estadística Aplicada i Mesura de la Sostenibilitat i el Desenvolupament

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE04. The ability to apply, critically and effectively, conceptual frameworks, data collection and processing techniques, applied statistics, mathematical modelling, systems analysis, geographic information systems, information and communication technologies and industrial ecology to meeting the challenges of sustainability and sustainable development.

TEACHING METHODOLOGY

MD1: Lecture or conference (EXP): Sharing knowledge through lectures by professors or by external guest speakers.

MD4: Tutorials of practical or theoretical works (TD): to perform an activity in the classroom, or a theoretical or practical exercise, individually or in small groups, with the advice of the teacher

and

MD6: Extensive project (PA): learning based in the design, planning and realisation in groups of a complex or extensive project or piece of work, applying and extending knowledge and writing a report on this approach and the results and conclusions

LEARNING OBJECTIVES OF THE SUBJECT

The main goal of this course is to provide a global view of the application of Data Science to real socio-environmental problem solving. The use of Data Mining techniques is presented in a complete Knowledge Discovery process devoted to extract relevant information from different kind of socio-environmental data (surveys, monitoring, data-warehouses...) to support decision-making from phenomena or organizations with high degrees of complexity. The course is focused to real socio-environemntal problems and to provide the proper elements to design efficient and correct Data Mining processes, according to the real problem targeted at every application, as well as to analyze the Data Scientist skills required to deal with. Main Data Mining methods are presented; training on several important practical aspects is provided, like effects on wrong pre-processing, wrong selection of data mining method, wrong interpretation of results or assumption of false hypothesis for the analyzed process; effective communication of results to decision-makers and reporting is also carefullly analyzed. This issues will help to guarantee the validity and utility of final results, as well as real impact of the analysis into the target domain. Real cases from socio-environmental field, like water management, sustainable touristic activities, pollution or land uses will be discussed to show the versatility of the discipline to provide better knowledge and decision suppport to a wide spectrum of very difficult real socio-environmental problems.

STUDY LOAD

Type	Hours	Percentage
Hours medium group	12,0	9.60
Hours large group	24,0	19.20
Self study	80,0	64.00
Hours small group	9,0	7.20

Total learning time: 125 h

CONTENTS

1. Introduction

Description:
1.1. Data Science, Data Mining, Knowledge Discovery from Databases and Intelligent decision support.
1.2. Data Mining Pillars: Statistics, Artificial Intelligence, Information Systems, Visualization

Specific objectives:
The Data Science and the overall process of Knowledge Discovery from Databases is presented, together with its steps and including Data Mining itself.
The disciplinary pillars of Data Mining are introduced: Statistics and Artificial Intelligence, Information Systems and Data Visualization
Finally, the basic schema of a Knowledge Discovery process is presented.

Related activities:
Presentation of the project to be developed along the course and working teams building

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

2. Scope, KDD process

Description:
2.1. Types of Problems suitable of Data Science
2.2. Ill-structured domains
2.3. A priori knowledge; Implicit knowledge. Causes and consequences
2.4. Main Data Mining Softwares (R, weka, rapid miner)

Specific objectives:
Different natures of real socio-environmental problems and their different levels of complexity are discussed according to the classification proposed by Simpson. Ill-structured domains are introduced, as well as a priori and implicit knowledge management, causes and consequences.
Some software tools for developing data mining tasks are presented, with special focus on R system.

Full-or-part-time: 2h
Theory classes: 2h

title english3. Formalising the Data Science problem and designing the complete Knowledge Discovery process

Description:
The steps of the Data Science process and the Knowledge Discovery process involved are introduced.

Related activities:
Define your project, identify data sources

Full-or-part-time: 1h
Theory classes: 1h

4. Data Structures

Description:
4.1 Main Socio-environmental data sources
4.2. Data and Metadata Representation

Specific objectives:
Main data structures analyzed by Data Mining techniques in socio-environmental fields.
Importance of metadata, formats and contents

Related activities:
Build the metadata file for your dataset

Full-or-part-time: 1h
Theory classes: 1h

5. Preprocessing

Description:
5.0 Reference preprocessing methodology
5.1. Data quality issues
5.2 Filtering and Sampling
5.3 Missing data treatment
5.4 Outliers
5.5 Data transformation and Derived data
5.6. Feature weighting and dimensionality reduction

Specific objectives:
Discussion on the importance of data quality and consequences of quality lack. Introduction of relevant aspects in data preparation step: Missing data, outliers detection and treatment, derived variables, transformed variables, filtering, sampling, feature weighting, dimensionality reduction (feature selection and factorial methods), all of them critical to guarantee the validity of the analysis. Good practice guidelines will be provided. Also a general reference methodology is provided

Related activities:
Preprocess your data for the project

Full-or-part-time: 5h
Theory classes: 5h

6. Choosing the proper Data Mining method

Description:
6.1. The problem-oriented approach
6.2 Criteria determining the suitability of a Data Mining method
6.3 The Data Mining Methods Conceptual Map (DMMCM-map)

Specific objectives:
The course follows a problem-oriented Data Science approach, where the nature of the problem mainly determines the analysis process and non vice-versa. Factors determining a correct choice of data mining method in real cases are discussed. The DMMCM typology of methods is presented as a conceptual basis for selection.

Related activities:
Designing the complete KDD process for your project and working plan

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

7. Data Mining Step: Descriptive Methods

Description:
7.1. Descriptive Methods
Clustering: partitioning methods, hierarchical, scalability. Hybrid methods, introduction of prior expert knowledge. Knowledge elicitation
7.2. Classes' characterization

Specific objectives:
Methods to identify and characterize profiles are presented

Related activities:
Cluster your data

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

8. Data Mining: Associative Methods

Description:
8.1. Association Rules mining
8.2 Factorial methods
8.3 bayesian networks

Specific objectives:
This chapter is devoted to methods discovering relationships between variables of the dataset

Related activities:
Use some associative method on your data

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

9. Data Mining: Discriminant Methods

Description:
9.1 Decision trees,
9.2 rule induction
9.3 support vector machines
9.4 discriminant analysis
9.5 Ensemble methods and bagging
9.6 hybrid methods.

Specific objectives:
Methods to predict a class variable (or a qualitative variable). At least 3 of them will be presented

Related activities:
Predict a qualitative variable by at least two discriminant methods

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

10. Data Míning: Predictive Methods

Description:
10.1 Regressión, statistical modelling in general.
10.2 Temporal methods
10.3 Artificial Neural Networks
10.4 Swarm Intelligence.

Specific objectives:
Methods to predict a numerical variable. At least 2 of them will be introduced

Related activities:
Predict (one or more) numerical variables

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

11. Spatio-temporal data mining

Description:
Spatio-temporal modelling

Specific objectives:
Some tools to deal with spatio-temporal data will be introduced

Related activities:
General review of project advances

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

12. Post-processing and validation

Description:
12.1. Post-processing tools
12.2. Model validation
12.3. Results validation

Specific objectives:
Post-processing tools and validation tools for both models and results adapted to different Data Mining methods.

Related activities:
Validation of models in your project.

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

13. Reporting and results communication

Description:
13.1 Reporting, automatic reporting
13.2 Results communication

Specific objectives:
Crucial to guarantee that the results of the Data Science process provide effective decision support to the end-user and the analysis have real impact on the target domain

Related activities:
Review of reporting the project

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

Scope of KDD process

Description:
concept and frontiers of the concept

Full-or-part-time: 2h
Theory classes: 2h

Data Science

Description:
Process, historical perspective, motivation and current impact
Design of a knowledge discovery process

Related activities:
Design the KDD process of practical work

Full-or-part-time: 2h
Theory classes: 2h

ACTIVITIES

Progress presentation of projects

Description:
Oral presentation of first part of project and discussion
Written deliverable

Specific objectives:
Milestone to sincronize all students with a suitable working plan
Communication and reporting skills are evaluated together with technical skills and organization of the working team

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

Final projects presentation

Description:
Oral presentation and written deliverable of the complete project. General and individual discussion with the teacher

Specific objectives:
Evaluation of the technical, communication and reporting skills, as well as the organizational performance of the working team

Full-or-part-time: 2h 30m
Theory classes: 2h 30m

GRADING SYSTEM

A long-term project will be developed by groups, applying a complete data science process to real data, including the application of methods lectured in the course. The project is developed under teachers? supervision.

An intermediate delivery (D1) will contribute to a better planning of the global work (D2). The final mark is assigned int he following way:

NotaFinal= 0.4D1+0.6D2

where

D1= 0.4*quality of written document +0.3* quality of oral presentation and discussion + 0.2* individual performance in laboratory sessions

D2= alfa * (0.4*quality of written document +0.3* quality of oral presentation and discussion + 0.2* individual performance in laboratory sessions)

being alfa a factor between 0.5 and 1.5 resulting from a cross-evaluation process done by the working team partners in D2 delivery.

AV2. Oral test to assess knowledge(PO).
AV3. Practical work developed individually or in groups along the course (TR). Includes the evaluation of results, reports, and oral presentation.
AV4. Attendance and participation in classes and laboratories (AP).
AV5. Quality and performance of team working (TG).