480093 - TDS - Socio-Environmental Data Science

Coordinating unit: 250 - ETSECCPB - Barcelona School of Civil Engineering
Teaching unit: 715 - EIO - Department of Statistics and Operations Research
Academic year: 2018
Degree: MASTER'S DEGREE IN SUSTAINABILITY SCIENCE AND TECHNOLOGY (Syllabus 2013). (Teaching unit Optional)
ECTS credits: 5

Teaching languages: English

Coordinator: KARINA GIBERT OLIVERAS

Others: Karina Gibert Oliveras
Miquel Sánchez-Marrè

Teaching staff

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

Opening hours

Timetable: Send an email to the corresponding lecturer to fix the meeting: karina.gibert@upc.edu, miquel@cs.upc.edu
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-environmental 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 carefully 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 support to a wide spectrum of very difficult real socio-environmental problems.

<table>
<thead>
<tr>
<th>Total learning time: 125h</th>
<th>Hours large group: 30h</th>
<th>24.00%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guided activities: 15h</td>
<td></td>
<td>12.00%</td>
</tr>
<tr>
<td>Self study: 80h</td>
<td></td>
<td>64.00%</td>
</tr>
</tbody>
</table>

Learning objectives of the subject
# Content

## 1. Introduction

<table>
<thead>
<tr>
<th>Description:</th>
<th>Learning time: 2h 30m</th>
</tr>
</thead>
</table>
| 1. Data Science, Data Mining, Knowledge Discovery from Databases and Intelligent decision support.  
1.2. Data Mining Pillars: Statistics, Artificial Intelligence, Information Systems, Visualization | Theory classes: 2h 30m |

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

| 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. | |

## 2. Scope, KDD process

<table>
<thead>
<tr>
<th>Description:</th>
<th>Learning time: 2h</th>
</tr>
</thead>
</table>
| 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) | Theory classes: 2h |

| 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. | |

## 3. Formalising the Data Science problem and designing the complete Knowledge Discovery process

<table>
<thead>
<tr>
<th>Description:</th>
<th>Learning time: 1h</th>
</tr>
</thead>
<tbody>
<tr>
<td>The steps of the Data Science process and the Knowledge Discovery process involved are introduced.</td>
<td>Theory classes: 1h</td>
</tr>
</tbody>
</table>

| Related activities: | |
|---------------------| |
| Define your project, identify data sources | |
# 4. Data Structures

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

**Related activities:**
Build the metadata file for your dataset

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

**Learning 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

**Related activities:**
Preprocess your data for the project

**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

**Learning time:**
5h  
Theory classes: 5h
## 6. Choosing the proper Data Mining method

**Learning time:** 2h 30m  
Theory classes: 2h 30m

**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)

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

**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.

## 7. Data Mining Step: Descriptive Methods

**Learning time:** 2h 30m  
Theory classes: 2h 30m

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

**Related activities:**
Cluster your data

**Specific objectives:**
Methods to identify and characterize profiles are presented
### 8. Data Mining: Associative Methods

**Learning time:** 2h 30m  
Theory classes: 2h 30m

**Description:**
- 8.1 Association Rules mining
- 8.2 Factorial methods
- 8.3 Bayesian networks

**Related activities:**
Use some associative method on your data

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

### 9. Data Mining: Discriminant Methods

**Learning time:** 2h 30m  
Theory classes: 2h 30m

**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

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

**Specific objectives:**
Methods to predict a class variable (or a qualitative variable). At least 3 of them will be presented
### 10. Data Mining: Predictive Methods

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

**Related activities:**
Predict (one or more) numerical variables

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

**Learning time:** 2h 30m
- Theory classes: 2h 30m

### 11. Spatio-temporal data mining

**Description:**
Spatio-temporal modelling

**Related activities:**
General review of project advances

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

**Learning 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

**Related activities:**
Validation of models in your project.

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

**Learning time:** 2h 30m
- Theory classes: 2h 30m
## 13. Reporting and results communication

**Learning time:** 2h 30m  
Theory classes: 2h 30m

**Description:**  
13.1 Reporting, automatic reporting  
13.2 Results communication

**Related activities:**  
Review of reporting the project

**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

### Planning of activities

<table>
<thead>
<tr>
<th>Event</th>
<th>Hours:</th>
<th>Description:</th>
<th>Specific objectives:</th>
</tr>
</thead>
</table>
| **Progress presentation of projects** | 2h 30m | Oral presentation of first part of project and discussion  
Written deliverable | Milestone to synchronize all students with a suitable working plan  
Communication and reporting skills are evaluated together with technical skills and organization of the working team |
| **Final projects presentation** | 2h 30m | Oral presentation and written deliverable of the complete project. General and individual discussion with the teacher | Evaluation of the technical, communication and reporting skills, as well as the organizational performance of the working team |
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 in the following way:

\[ \text{NotaFinal} = 0.4D1 + 0.6D2 \]

where

\[ D1 = 0.4 \times \text{quality of written document} + 0.3 \times \text{quality of oral presentation and discussion} + 0.2 \times \text{individual performance in laboratory sessions} \]

\[ D2 = \alpha \times (0.4 \times \text{quality of written document} + 0.3 \times \text{quality of oral presentation and discussion} + 0.2 \times \text{individual performance in laboratory sessions}) \]

being \( \alpha \) 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).