240615 - An Introduction to Data Science

Coordinating unit: 240 - ETSEIB - Barcelona School of Industrial Engineering
Teaching unit: 715 - EIO - Department of Statistics and Operations Research
Academic year: 2018
Degree: BACHELOR’S DEGREE IN INDUSTRIAL TECHNOLOGY ENGINEERING (Syllabus 2010). (Teaching unit Optional)
BACHELOR’S DEGREE IN MATERIALS ENGINEERING (Syllabus 2010). (Teaching unit Optional)
BACHELOR’S DEGREE IN CHEMICAL ENGINEERING (Syllabus 2010). (Teaching unit Optional)
ECTS credits: 4.5  Teaching languages: English

Teaching staff
Coordinator: JOSEP GINEBRA
Others: JOSEP GINEBRA

Opening hours
Timetable: Monday and Wednesday from 11:30 to 13:30

Prior skills
To have passed Estadística.

Degree competences to which the subject contributes

Specific:
1. Basic knowledge on the use and programming of computers, operative systems, data bases and computer software with an engineering application.
2. Knowledge and capacities to organise and manage projects. Knowing the organisational structure and functions of a project office.

Transversal:
3. EFFICIENT ORAL AND WRITTEN COMMUNICATION. Communicating verbally and in writing about learning outcomes, thought-building and decision-making. Taking part in debates about issues related to the own field of specialization.
4. TEAMWORK. Being able to work as a team player, either as a member or as a leader. Contributing to projects pragmatically and responsibly, by reaching commitments in accordance to the resources that are available.

Teaching methodology
All classes will be taught in a computer lab. The data analysis will be done with MINITAB and with R. Every week there will be small data analysis assignments to be done at home. Students will have to do a final project.

Learning objectives of the subject
At the end of the course the student should be able to identify situations where it is useful to analyze data, to identify the model and/or method of analysis that is best for his data, to build a model that summarizes the information in the data and allows to make predictions, to reduce the dimensionality and visualize multivariate data, to implement supervised and unsupervised classification algorithms, and to evaluate the quality of the results obtained.
# Study load

| Total learning time: 112h 30m | Hours large group: 0h 0.00% | Hours medium group: 45h 40.00% | Hours small group: 0h 0.00% | Guided activities: 0h 0.00% | Self study: 67h 30m 60.00% |
# 240615 - An Introduction to Data Science

## Content

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Description</th>
<th>Learning time</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chapter 1: Introduction</strong></td>
<td>(ENG) 1.- Problems: Association, prediction and classification. 2.- Tools: Statistical models and multivariate analysis.</td>
<td>3h 30m</td>
</tr>
<tr>
<td><strong>Chapter 2: Linear models for continuous response</strong></td>
<td>(ENG) 1.- Normal linear model. 3.- Model fit; least squares and robust regression. 3.- ANOVA table and goodness of fit measures. 4.- Inference on the model parameters. 5.- Prediction. 6.- Model checking. 7.- Model selection. 8.- Cross validation and lack of fit tests. 9.- Model interpretation; Bias, collinearity and causality. 10.- Use of categorical explanatory variables. 11.- Comparison of means. 12.- Analysis of two-level factorial designs.</td>
<td>30h</td>
</tr>
<tr>
<td><strong>Chapter 3: Non-linear models for a continuous response</strong></td>
<td>1.- Normal non-linear model. 2.- Model fit. 3.- Inference. 4.- Model checking.</td>
<td>6h</td>
</tr>
<tr>
<td><strong>Chapter 4: Categorical and discrete response models</strong></td>
<td>(ENG) 1 Generalized linear model. 2. - Count response models. 3.- Binary response models. 4.- Model fit. 5.- Inference. 6.- Model checking. 7.- Prediction. 8.- Model interpretation. 9.- Contingency tables and models for a polytomous response.</td>
<td>22h 30m</td>
</tr>
</tbody>
</table>
### Chapter 5: Time series models

**Learning time:** 13h  
Theory classes: 3h  
Laboratory classes: 3h  
Guided activities: 3h  
Self study: 4h

**Description:**  
1. Description of a time series.  
2. AR models.  
3. MA models.  
4. ARIMA models.  
5. Seasonal ARIMA models.

### Chapter 6: Visualization of multivariate data (Dimensionality reduction)

**Learning time:** 6h  
Theory classes: 1h 30m  
Laboratory classes: 1h 30m  
Self study: 3h

**Description:**  
(ENG) 1. Principal components analysis.  
2. Correspondence analysis.

### Chapter 7: Cluster analysis (Unsupervised classification)

**Learning time:** 6h 30m  
Theory classes: 1h 30m  
Laboratory classes: 1h 30m  
Guided activities: 1h 30m  
Self study: 2h

**Description:**  
1. Hierarchical methods.  
2. Partition methods (k-means algorithm).  
3. Variable cluster analysis.

### Chapter 8: Discriminant analysis (Supervised classification)

**Learning time:** 8h 30m  
Theory classes: 1h 30m  
Laboratory classes: 1h 30m  
Guided activities: 1h 30m  
Self study: 4h

**Description:**  
1. Linear discriminant.  
2. Quadratic discriminant.  
3. Logistic discriminant.
Chapter 9: Non-parametric regression and classification models

Description:
1.- Local smoothers. 2.- Nearest neighbors. 3.- Additive models. 4.- Classification and regression trees. 5.- Neural networks.

Learning time: 4h 30m
Theory classes: 1h 30m
Laboratory classes: 1h 30m
Self study: 1h 30m

Qualification system

There will be a take home midterm exam and an in class final exam.

Grade = 0.1 Assignments + 0.3 Final Project + 0.1 Midterm + 0.5 Final Exam

Bibliography

Basic:


Complementary:


