Course guides

240615 - 240615 - An Introduction to Data Science

Unit in charge: Barcelona School of Industrial Engineering
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.
Degree: BACHELOR’S DEGREE IN INDUSTRIAL TECHNOLOGY ENGINEERING (Syllabus 2010). (Optional subject).
BACHELOR’S DEGREE IN MATERIALS ENGINEERING (Syllabus 2010). (Optional subject).
BACHELOR’S DEGREE IN CHEMICAL ENGINEERING (Syllabus 2010). (Optional subject).

Academic year: 2020 ECTS Credits: 4.5 Languages: English

LEKTURER

Coordinating lecturer: JOSEP GINEBRA
Others: JOSEP GINEBRA

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 assignements to be done at home. Students will have to do a final project. On the Q1 of 2020/21 the classes will be online, but students will also have the option to attend the class in person if they want to.

LEARNING OBJECTIVES OF THE SUBJECT

At the end of the course the student should be able to identify situations where it is useful to analize 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

<table>
<thead>
<tr>
<th>Type</th>
<th>Hours</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self study</td>
<td>67.5</td>
<td>60.00</td>
</tr>
<tr>
<td>Hours medium group</td>
<td>45.0</td>
<td>40.00</td>
</tr>
</tbody>
</table>

Total learning time: 112.5 h

CONTENTS

Chapter 1: Introduction

Description:
(ENG) 1.- Problems: Association and prediction. 2.- Tools: Statistical models and multivariate analysis.

Full-or-part-time: 3h 30m
Theory classes: 1h 30m
Guided activities: 1h
Self study: 1h

Chapter 2: Linear models

Description:
(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 and cross validation. 8.- Model interpretation; Bias, collinearity and causality. 9.- Use of categorical explanatory variables. 10.- Comparison of means.

Full-or-part-time: 30h
Theory classes: 6h
Laboratory classes: 6h
Guided activities: 6h
Self study: 12h

Chapter 3: Non-linear models

Description:
1.- Normal non-linear model. 2.- Model fit. 3.- Inference. 4.- Model checking.

Full-or-part-time: 6h
Theory classes: 1h 30m
Laboratory classes: 1h 30m
Self study: 3h
Chapter 4: Categorical and count response models

Description:
(ENG) 1.- Binary logistic model. 2.- Model fit. 3.- Inference. 4.- Model checking. 5.- Prediction. 6.- Model interpretation. 7.- Nominal logistic model. 8.- Contingency tables and logistic model. 9.- Count response model.

Full-or-part-time: 22h 30m
Theory classes: 4h 30m
Laboratory classes: 4h 30m
Guided activities: 4h 30m
Self study : 9h

Chapter 5: Time series models

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

Full-or-part-time: 13h
Theory classes: 3h
Laboratory classes: 3h
Guided activities: 3h
Self study : 4h

Chapter 6: Visualization of multivariate data (Dimensionality reduction)

Description:
(ENG) 1.- Introduction. 2.- Principal components analysis. 3.- Correspondence analysis.

Full-or-part-time: 6h
Theory classes: 1h 30m
Laboratory classes: 1h 30m
Self study : 3h

Chapter 7: Cluster analysis (Unsupervised classification)

Description:
1.- Introduction. 2.- Hierarchical methods. 3.- Partition methods (k-means algorithm). 4.- Variable cluster analysis.

Full-or-part-time: 6h 30m
Theory classes: 1h 30m
Laboratory classes: 1h 30m
Guided activities: 1h 30m
Self study : 2h

Chapter 8: Discriminant analysis (Supervised classification)

Description:
1.- Introduction. 2.- Linear discriminant. 3.- Quadratic discriminant. 4.- Logistic discriminant.

Full-or-part-time: 8h 30m
Theory classes: 1h 30m
Laboratory classes: 1h 30m
Guided activities: 1h 30m
Self study : 4h
Chapter 9: Non-parametric regression and classification models

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

Full-or-part-time: 4h 30m
Theory classes: 1h 30m
Laboratory classes: 1h 30m
Self study: 1h 30m

GRADING 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

During the 2019-20 spring semester, as a consequence of the covid19 crisis, the qualification method will be the same one, with the only difference that the final exam will not be in class, but a take home exam.

BIBLIOGRAPHY

Basic:

Complementary: