# Big Data for Business Intelligence

**Academic Year: ( 2018 / 2019 )** **Review date: 07-05-2018**

**Department assigned to the subject: Statistics Department**

**Coordinating teacher: AUSIN OLIVERA, MARIA CONCEPCION**

**Type: Electives ECTS Credits : 6.0**

**Year : 4 Semester :**

## REQUIREMENTS (SUBJECTS THAT ARE ASSUMED TO BE KNOWN)

Statistics I
Statistics II
Data analytical techniques for business
Introduction to data mining for business intelligence

## OBJECTIVES

1. Understand the importance of transforming large volumes of data into relevant information for decision making and business development in organizations, companies and individuals.

2. Learn the basic techniques of preprocessing and visualization of data. Gain knowledge on methods to work with missing and atypical data. Acquire the ability to use of dimension reduction techniques.

3. Gain knowledge on the main methods of supervised learning in regression and their usefulness in prediction problems. Distinguish between linear and non-linear models and understand the importance of model selection methods.

4. Become familiar with the usual supervised learning procedures for classification. Understand the most common classifiers and their limitations. Gain knowledge in advanced methods for classification and their benefits in business.

5. Be able to identify the appropriate Big Data techniques in real business problems: customer classification, scoring, risk management, fraud detection, bankruptcy prediction, etc.

## DESCRIPTION OF CONTENTS: PROGRAMME

1. Introduction.

2. Data collection, sampling and preprocessing.
2.1. Types of data.
2.2. Sampling.
2.3. Data visualization tools.
2.4. Missing values.
2.5. Outlier detection and treatment.
2.6. Data transformations.
2.7. Dimension reduction.
2.8. Application: Risk management in the stock market.

3. Supervised learning: regression.
3.1. Linear and polynomial regression.
3.2. Cross-validation.
3.3. Model selection and regularization methods (ridge and lasso).
3.4. Nonlinear models, splines and generalized additive models.
3.5. Application: credit-scoring prediction.

4. Supervised learning: classification.
4.1. Bayes classifiers
4.2. Logistic regression.
4.3. K-nearest neighbors.
4.4. Random forest.
4.5. Support-vector machines.

4.6. Boosting.
4.7. Application: Credit risk.
4.8. Application: Fraud detection.
4.9. Application: Bankruptcy prediction

## LEARNING ACTIVITIES AND METHODOLOGY

Theory (2 ECTS). Lectures with available material posted in internet. Problems (4 ECTS) Problem Solving classes. Computational exercises at computer room. Work assignments in groups. Weekly office hours to assist students on an individual and group basis.

## ASSESSMENT SYSTEM

Written test (50%). Exercises and computational work (5%). Work in class (10%). Group project (10%) Quiz 1 (15%). Group quiz (10%). A minimum mark of 4 points out of 10 is required.

| | |
|---|---|
| **% end-of-term-examination:** | 60 |
| **% of continuous assessment (assigments, laboratory, practicals…):** | 40 |

## BASIC BIBLIOGRAPHY


- Bradley Efron, Trevor Hastie. Computer Age Statistical Inference: Algorithms, Evidence and Data Science., Cambridge University Press, 2016
- E. Alpaydin Introduction to Machine Learning, MIT Press., 2010
- James, G., Witten, D., Hastie, T., Tibshirani, R. An Introduction to Statistical Learning with Applications in R, Springer, 2013.
- T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer. , 2009
- Trevor Hastie, Robert Tibshirani, Martin Wainwright Statistical Learning with Sparsity: the Lasso and Generalizations, Chapman & Hall, 2015