# **Applied statistics and big data**

Module A: Prof. Dimitris Fouskakis; Module B: Prof. Marco De Ieso

***COURSE AIMS AND INTENDED LEARNING OUTCOMES***

Firms work with data sets containing a large number of observations and many variables. The aim of the course is to provide students with the statistical methods for dealing with data of this kind. It will provide methods, models and tools to analyze and interpret empirical data, with special emphasis on economic and business problems. The course is divided into two modules.

In Module A students will learn multivariate data analysis techniques to perform data analyses. The study will be carried out using the powerful open-source *R language,* which is rapidly becoming the most popular environment for statistical computing and graphics. Students will be given a hands-on introduction to *R,* including practical lab sessions. The following learning abilities are provided and expected to be achieved by participants at the end of the module: (a) Knowledge of concepts, terms and methods of descriptive statistics, statistical inference and probability; (b) Basic knowledge of the statistical programming language *R*; (c) Ability to apply knowledge and understanding the statistical methods required to analyze categorical data and to draw conclusions from such methods; (d) Ability to apply knowledge and understanding multiple linear regression models, by correctly interpreting the output, checking the assumptions and performing predictions; (e) Ability to apply knowledge and understanding multiple logistic regression models, for binary responses, and correctly interpreting the results in terms of odd ratios.

In Module B students will learn methods of supervised and unsupervised machine learning. The study will be carried out using the powerful open-source *R language.* Students will be given lectures about Clustering, Principal Component Analysis and Decision Trees Algorithm. Practical lab sessions are included. The following learning abilities are provided and expected to be achieved by participants at the end of the module: (a) Knowledge of concepts, terms and methods of supervised and unsupervised learning; (b) Basic knowledge of libraries used to apply Clustering, Classification and Principal Components Analysis Algorithms; (c) Ability to apply knowledge and understanding Classification models; (d) Ability to apply knowledge and understanding Cluster models; (e) Ability to apply knowledge and understanding to reduce the dimensionality of data.

***COURSE CONTENTS***

Module A: *Methods*

Basic Notions of Descriptive Statistics, Statistical Inference and Probability. Introduction to the R language including graphs, probability distributions and simulation. Contingency tables for analyzing categorical data and hypotheses tests. Multiple linear regression: output interpretation, categorical explanatory variables, regression diagnostics, model selection, predictions. Logistic regression models for binary responses and interpretation of the output in terms of odds ratios.

Module B: *Applications*

Development of a data mining study: The CRISP-DM methodology. Business Understanding and Data Understanding: getting to know your data. Data Preparation: getting your data ready for analysis. Principal Components Analysis: Computation of the Principal Components, Interpretation of the Principal Component Analysis. Cluster Analysis: Affinity Measures, Single / Complete / Average linkage Algorithms, K-Means Algorithm, output interpretation. Classification and scoring: Quality measures, the C4.5 method, Evaluation of classification models.

***READING LIST***

In Module A notes will be provided via Blackboard.

In Module B notes will be provided via Blackboard.

***TEACHING METHOD***

Module A involves lectures and PC-labs using the R language, requires active participation, and ongoing personal study.

Module B involves lectures and PC-labs using R language, require active participation, and ongoing personal study.

***ASSESSMENT METHOD AND CRITERIA***

First partial exam: written take-home exam with open-ended questions and exercises on topics in Module A, to be handed in within 10 days.

Second partial exam: in-class written exam with open-ended questions and exercises on topics in Module B.

The general exam consists of both

– A written take-home exam to be handed in within 10 days, on Module A. Open-ended questions and exercises.

– A written in-class exam on Module B. Open-ended questions and exercises.

The second partial exam will take place together with the in-class part of the first general exam.

Separate grades for the take-home and in-class exam will be provided. The final grade will be the average of the two.

***NOTES AND PREREQUISITES***

Correspondence between students and teachers will be managed through Blackboard. The course is a natural follow-up to the first Statistics course. Thus, students enrolling in this course should have a basic knowledge of Probability and Statistics.