# Computer Systems Laboratory for Decision-Making Support (Data Mining)

## I Mod.: Prof. Gabriele Cantaluppi; II Mod.: Prof. Stefano Gatti

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

The aim of the course is to teach students the main concepts, terms and methods of statistical machine learning and the ability to apply them correctly to problems arising in economics and management. The course will provide also instruments to let students master rigorous reasoning and data-driven decision-making, useful for quantitative analyses in other courses of the curriculum, as well as for analyses required in careers in all fields involving data management.

Particular attention will be paid to methodological and practical aspects, through illustration of real cases and classroom exercises with the aid of software tools widely used in the academic and corporate world.

Upon successful completion of the course the student will be able to master the following skills and knowledge:

1. knowledge of the concepts, terms and methods of statistical machine learning and of the principal functions of Microsoft Excel and Python software with Jupyter Notebook (DD1- Knowledge and understanding);
2. ability to correctly apply statistical and machine learning methods to problems in economics and business management (DD2- Applying knowledge and understanding);
3. quantitative thinking addressed to make independent judgements, driven by descriptive and inferential statistics statements (DD3- Making judgements);
4. ability to build and read output from statistical and machine learning procedures presented in the course and draw conclusions from them, by means of the extraction of qualitative information from data, with clarity and accuracy and in forms that are suitable for the audiences being addressed, both orally and in writing (DD4-Communication);
5. mastery of statistical and machine learning methods, rigorous reasoning and data-driven decision-making, useful for quantitative analyses in other courses of the curriculum, as well as for analyses required in careers in all fields involving data management (DD5- Lifelong learning skills).

***COURSE CONTENTS***

The course includes 60 hours of lessons and 8 of laboratories (equally divided between the two modules).

– In the first module (statistical theory, 30 hours), which allows in-depth study of aspects concerning the application of statistical methodologies for the analysis of business data, special attention is paid to the procedures available in Excel. These include multiple regression analysis, simulation of probability distributions and time series analysis (moving averages, seasonality identification and adjustment methods, exponential smoothing and scenario forecasting).

– The second module (computer science, 30 hours) is devoted to in-depth practical analysis of business data and application of data mining methods (classification, clustering, association rules) aimed at addressing typical business problems (profiling, propensity analysis, segmentation, market basket analysis).

***READING LIST***

For the statistics part

G. Boari, G. Cantaluppi, *Raccolta di temi ed esercizi per il corso di Laboratorio Informatico per le Decisioni Aziendali,* 2° modulo (statistico), EDUCatt, Milano, 2017.

Recommended texts on statistical inference and data mining

G. Cicchitelli-P. D’Urso-M. Minozzo  *Statistica: principi e metodi,* Pearson, Milano, 2018.

G. James-D. Witten-T. Hastie-R. Tibshirani, *An Introduction to Statistical Learning,* Springer, New York, 2013, https://www.statlearning.com/ (ch. 1, 2, 3, 5).

For the computer science part

Jupyter Team, *Jupyter Project Documentation,* https://docs.jupyter.org/en/latest/index.html.

J. VanderPlas, *Python Datascience handbook,* O’Reilly Media Inc, USA, 2016, https://jakevdp.github.io/PythonDataScienceHandbook/.

Recommended texts for in-depth study for the theoretical models and business decision making.

M. Bramer, *Principles of Data Mining,* Springer, 2020, 4th Edition.

C. Vercellis, *Business Intelligence: modelli matematici e sistemi per le decisioni,* McGraw-Hill, Milano, 2006.

***TEACHING METHOD***

Theorethical lessons (equally divided between theoretical framing of topics and development of computer skills) and self-learning tutorials.

***ASSESSMENT METHOD AND CRITERIA***

Written exam in order to test the whole program at the end of the lectures. A midterm exam will take place at the end of the first module.

For the statistics part, the exam consists of three exercises, inclusive of open-ended questions. Two exercises are of numerical type, one is theoretical.

For the second module (computer science), the test includes open-ended, closed-ended questions and a practical data analysis exercise in which the student must write a small program in Python using Jupyter Notebook.

Evidence from the two modules carries equal weight with reference to the final assessment. In order to pass the exam, the student must achieve a passing grade in both modules.

The exam is designed to assess first and foremost reasoning skills and analytical rigor on the topics covered in the course, as well as language properties and communication skill.

***NOTES AND PREREQUISITES***

Due to the structure of class activities, students are recommended to use their notebook or netbook during theoretical lessons, which take place in traditional rooms.

The following software are adopted in the course:

1. Microsoft Excel

2. Jupyter Notebook, open source, available for download at: https://jupyter.org/.

3. Python, open source, available for download at: https://www.python.org/.

Students enrolling in this course should have a basic understanding of statistics with regard to data analysis, probability and inference, at the level of the combined courses ‘Statistica I’/‘Statistica (Analisi dei dati e probabilità)’/Statistics and ‘Statistica applicata’ taught in this University. Introductory lectures will be devoted to review some subjects presented in the course ‘Statistica applicata’.