# Applied Statistics and Big Data Analytics

Dr. Mario Barbato

## COURSE AIMS AND INTENDED LEARNING OUTCOMES

This course aims at introducing the students to the informatics and statistical tools involved in the analysis of high-dimensional data (Big Data). The course will reinforce students’ previous knowledge in statistics, and expand it with a focus on multivariate statistics, while providing the informatics tools to translate the theory into practice. The most common Machine Learning approaches, their requirements, applications and limitations, will be introduced along with practical examples on the fields of Viticulture and Enology.

On completion of this course, the students will be able to:

* Use informatics to extract, explore and manipulate information from low- and high-dimensional datasets.
* Formulate precise questions, and identify the best statistical tool to test the hypotheses.
* Identify the most suited framework for a given hypothesis-testing scenario.
* Build a coherent and functional analysis pipeline including bespoke functions when appropriate.
* Be able to complement the results with the most appropriate graphical representations.

## COURSE CONTENT

|  |  |
| --- | --- |
| **Topics** | **CFU** |
| Introduction to programming:* Introduction to R programming.
* Data types and data editing.
* Algorithm thinking and main constructs.
* Plots, graphical representations, and reports in R.
 | 1.25 |
| Catch-up on statistics:* Data distributions and descriptors.
* Hypothesis testing, data simulation and resampling techniques.
* Statistical models, ANOVA, GLM and GLMM.
 | 1.25 |
| High-dimensional data:* Data collection, homogeneity, and conversion.
* Data clustering: centroid- and density-based clustering, PCA.
* Algorithm building: efficiency and scalability.
 | 1.25 |
| Algorithms for high-dimensional data and Statistical Learning:* Introduction on supervised, unsupervised and reinforcement learning.
* Classification problems: PLS-DA and Random Forest.
* Introduction on Neural Networks and deep-learning.
 | 1.25 |
| **Practical** |  |
| * Data browsing and editing.
* Sensors data analysis: heterogeneous data type in viticulture.
* Classification using complex food matrix data.
 | 1.00 |

## READING LIST

There is no specific course textbook. Reading material, references, presentations, and code will be made available at the end of each section of the course.

## TEACHING METHOD

The teaching method will combine lectures and practical activities. Lectures will cover the theoretical topics of the course, while providing a set of real-life examples relevant to Viticulture and Enology. The practical activities will aim at reinforcing the theoretical information and transform them in challenges first, and practical skills then.

## ASSESSMENT METHOD AND CRITERIA

At the end of the course the students will be graded following two written tests (total maximum score is 30/30). The first test will consist in 20 multiple-choice questions (maximum score is 20 points, then translated in 30/30); the students will have 20 minutes to complete the test. Each question correctly answered (in full) will award 1 point, incorrect or partially-correct answers will be awarded proportinally less points. The second test will consist in developing an analisys pipeline (maximum score is 30/30); the students will have two hours to complete it. The resulting report will be evaluated considering the rationale proposed and its implementation.

The average mark between quiz and pipeline will reflect the final score. Students wishing to improve the final mark can take the oral exam.

## NOTES AND PREREQUISITES

* Knowledge of elementary statistics is required.
* Lectures will be delivered through frontal teaching. If the current health emergency prevents frontal teaching, remote teaching can be achieved through synchronous or asynchronous lectures.
* Dr. Mario Barbato is available to meet students upon request at the Department of Animal Science, Food and Nutrition (DIANA).