# IT coding for data science (first module 6 cfu; second module 6 cfu)

## Prof. Enrico Barbierato

### **FIRST MODULE (6 CFU)**

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

The course aims at providing strong programming skills, presenting, at the same time, a general introduction to Computer Science. At the end of course, the students will be able to solve problems of average complexity by using the fundamental algorithms and developing scripts in R and Python scripting languages.

This class corresponds to the first module of the course IT Coding for Data Science

***COURSE CONTENT***

*Computer architectures:* CPU, Bus, Registers. Main memory: RAM e ROM. Secondary memories: magnetic and optical units. Information coding: ASCII 7 and 8 bits. Floating-point representation and binary arithmetic. Boolean algebra and gates.

*Abstract data structures:* array, matrix, list, set, dictionary, dataframes, trees and graphs.

*Algorithms theory:* fundamental constructs (assignment, test, loop). Elementary problems. Notable algorithms: sequential search, binary and hash. Sorting algorithms: bubble. Complexity. Recursive functions.

*Python language:* Preliminaries on Pycharm and Jupyter. Elementary data types and abstract data structures. Testing conditions and loops. Custom functions. Numpy library. Dataframes and Pandas library. Classes and objects. Encapsulation, inheritance and polimorphism. Magic methods. Graphical functions. GUI (Tkinter)

*R language:* Preliminaries on RStudio. Elementary data types and abstract data structures. Testing conditions and loops. Custom functions. Dataframes. Graphical functions. I/O: Writing and reading a text file locally and remotely. CSV files.

*Introduction to relational database*: Relational databases, relational algebra, simple SQL statements.

***READING LIST***

Powerpoint slides. Scientific articles.

***TEACHING METHOD***

The course will include lectures and class exercises based on traditional teaching and specific examples. It is strongly advised to attend to lectures for working on case studies and examples, and for revising materials.

The course also involves lectures and exercise sessions using the integrated development environments Rstudio, PyCharm and Jupyter. Active participation, and ongoing personal study are required.

***ASSESSMENT METHOD AND CRITERIA***

The exam consists of the following weighted parts:

80% Written test, including open or multiple choice questions and programming exercises;

10% Individual small project

10% Discussion of a paper from the scientific literature in the area of Computer Science.

***NOTES AND PREREQUISITES***

***Prerequisites:*** Logical thought and strong organizational skills are a must. Previous experience with programming is a plus although not mandatory.

Attendance is strongly recommended.

***Day and reception hours:*** Students can contact the instructor by e-mail to arrange for day and reception hours at the instructor’s workplace.

In case the current Covid-19 health emergency does not allow frontal teaching, remote teaching will be carried out following procedures that will be promptly notified to students.

### **SECOND MODULE (6 CFU)**

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

The course aims at providing solid programming skills, presenting, at the same time, a general introduction to Data Science. At the end of course, the students will be able to clean, analyze, and finally visualize data by using R and Python scripting languages and their most important packages. Furthermore, the students will be able to put in practice Data Science techniques.

This class corresponds to the second module of the course IT Coding for Data Science. 2

***COURSE CONTENT***

The following topics will be taught by using R and Python programming language, scientific and visualization libraries, such as SciPy, NumPy, Matplotlib, Sklearn.

*Exploratory Data Analysis (EDA)*: Variable identification (continuous and categorical). Univariate analysis: box plots, histograms, bar charts. Bi-variable analysis - determine the interaction between variables by building visualization tools: scatter plots, stacked column chart, boxplots. Detect and treat missing values and outliers

*Math for Machine Learning*: Space. Vectors and matrices. Math operators. Dot product and projection. Determinant and inverse of a matrix. Eigenvalues and Eigenvectors. Derivatives and partial derivatives of a function.

*Data Science*: Applications of eigenvectors: Google page ranking algorithm and PCA. Clustering. Regression. Classification. Metrics.

*Statistics for Data Science:* Random variables. Probability distribution functions (PDFs). Mean, Variance, Standard Deviation. Covariance and Correlation . Linear Regression and Ordinary Least Squares (OLS). Parameter properties (Bias, Consistency, Efficiency). Confidence intervals. Hypothesis testing. Statistical significance. Type I & Type II Errors. Statistical tests (Student's t-test, F-tests). p-value.

*Advanced Python programming:* Synthetic data generation. Pipelines. Iterators. Lambda functions.

- *Advanced Topics*: Time series analysis. Discrete Time Markov Chains. Recommendation systems.

***READING LIST***

Powerpoint slides and articles

***TEACHING METHOD***

The course will include lectures and class exercises based on traditional teaching and specific examples. It is strongly advised to attend to lectures for working on case studies and examples, and for revising materials.

The course also involves lectures and exercise sessions using the integrated development environments RStudio, PyCharm and Jupyter. Active participation, and ongoing personal study are required.

***ASSESSMENT METHOD AND CRITERIA***

The exam consists of the following weighted parts:

60% Written test, including open or multiple choice questions, manual exercises and Python/R coding;

30% Individual small project

10% Discussion of a paper from the scientific literature in the area of Computer Science.

***NOTES AND PREREQUISITES***

***Prerequisites (recommended although not mandatory)****:*

* *Probability and Statistics course*
* *Data Analysis, Techniques and tools*

***Day and reception hours:***

Students can contact the instructor by e-mail to arrange for day and reception hours at the instructor’s workplace.