# 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 scripting 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

*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.

*Microsoft Excel:* Basic operations. Pivot Tables. Graphics. Working with Datasets

*Explorative analysis:* data visualization. Statistical measures, quartiles and boxplots. Missing values and outliers. Univariate, bivariate and multivariate analysis.

*Statistical inference:* probability. Random variables. Probability mass and density functions. Distributions and Asymptotics. Confidence intervals. Hypothesys testing and P-value.

***READING LIST***

Powerpoint slides.

***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 grade is calculated as follows:

10% Individual assignments;

40% Written exam, including open questions and exercises;

25% Individual project

25% Discussion of a paper from the scientific literature.

***NOTES AND PREREQUISITES***

***Prerequisites:*** 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 strong 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.

***COURSE CONTENT***

Each one of the following topics will be taught by using Python programming language, scientific and visualization libraries, such as SciPy, NumPy, Matplotlib, Sklearn.

*Data science*: Introduction. What is Data Science about? Data Science and AI.

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

*Introduction to Machine Learning*: Applications of eigenvectors: Google page ranking algorithm and PCA. Clustering: K-means, C-Means and DBSCAN. Hierarchical clustering. Linear Regression, Metrics (Jaccard, R2 and F1). Multilinear, Non linear, Logistic Regression, K-Nearest Neighbors, Naive Bayes, Support Vector Machines. Associative rules, Decision Trees. Random Forests

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

*Introduction to Big Data*: The Hadoop architecture. Apache Spark. SQL for Data Science. Not relational DBs (Mongo). Data warehouse and Data Lake.

*Introduction to Advanced Topics*: Multitasking in Python and GPU. Data Fairness. Advanced AI: Explainable AI (XAI) with Lime Python library. Recommendation Systems. Time series analysis.

*Notable Tools*: Microsoft BI (if time allows, one of *Tableau* or *AutoML*).

***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 grade is calculated as follows:

25% Assignments;

25% Individual project

50% Written exam, including exercises;

***NOTES AND PREREQUISITES***

***Prerequisites:*** Basic knowledge of elementary skills on computer programming is requested.

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.