IProML: Introduction to Programming and Machine Learning in Python, 2022

Syllabus

Course responsible:

Andrea Vandin, andrea.vandin@santannapisa.it https://www.santannapisa.it/en/andrea-vandin
Daniele Licari, daniele.licari@santannapisa.it https://www.linkedin.com/in/daniele-licari/

Language: English

Duration:
Module 1 16h, From 20/04/2022 to 16/05/2022
Module 2 14h, From 20/05/2022 to 13/06/2022

Description:
The course introduces students to programming, data analytics and the basis of machine learning, using python as a reference language.

- Module 1 introduces students to the fundamental principles of 'good' programming with basic applications to data processing. The module starts from basic notions of programming (variables, data types, collections, control & repetition structures, functions & modules, object-oriented programming), up to medium-complex data processing functionalities (loading, manipulation, and visualization of CSV/online data).
- Module 2 introduces the students to the components of typical data analysis processes and machine learning pipelines. It first builds the necessary toolset by introducing popular Python libraries for data manipulation/visualization (NumPy, Pandas, Seaborn, scikit-learn), applied to simple applications. The toolset is then applied to a more complex case study on the classification of benign and malignant breast cancer, including aspects of data preprocessing, dimensionality reduction, clustering, and classification. The module concludes by presenting KNIME, a popular python-integrated workflow-based language for data analysis.

A student who has met the objectives of the course will acquire an understanding of the issues and tasks involved in 'good' computer programming and data analysis, to be able to make informed decisions. The student will be able to write python programs of various nature, with a focus on complex data analysis and predictive tasks.

Prerequisites: No prerequisites for Module 1, while Module 2 requires knowledge of computer programming (possibly obtained attending Module 1).

Materials:
The course makes extensive use of online repositories and game-based e-learning platforms to

- GitHub Wiki (website): collect slides, coding examples, datasets, and further course material
- Colab: distribute and automatically provide feedback for weekly coding assignments
- Kahoot: perform online quizzes to monitor the learning process

Suggested books are:

- Learning Python, M. Lutz
- Python for Data Analysis, W. McKinney
- Statistics and Machine Learning in Python, E.Duchesnay, T.Löfstedt, F.Younes

We will use **python** as the programming language and statistical software of choice for the course.

Evaluation:
Students can attend single modules. These are 'attività trasversali', hence there will not be an exam, but an attendance certificate (attestazione di presenza) with mandatory attendance of at least 80%.

Attendance:
The course will likely be conducted remotely.

Tentative Schedule:

Module 1 – 16 hours

Class	Topic	Date	Time
1	Course Introduction Console I/O & Variables	Wed 20/04	17:00-19:00
2	Data types & Operations	Fri 22/04	17:00-19:00
3	Collections First plots *Practicum*	Wed 27/04	17:00-20:00
4	Control statements CSV manipulation on COVID19 data *Practicum*	Mon 02/05	17:00-20:00
5	Functions Creation of word clouds from online news	Fri 06/05	17:00-19:00
6	Modules Exceptions OOP	Mon 09/05	17:00-19:00
7	Tutorial on Process-oriented Data Science: Process mining	Mon 16/05	17:00-19:00

Module 2 – 14 hours

Class	Topic	Date	Time
1	Course & Project introduction Advanced Python IDEs (JuPyteR, Google Colab) Advanced libraries for data manipulation: NumPy	Fri 20/05	17:00-19:00
2	Advanced libraries for data manipulation: Pandas 1 Application to official Italian COVID'19 data Application to Yahoo! Finance stock prices	Mon 23/05	17:00-19:00
3	Advanced libraries for data manipulation: Pandas 2 Application to Yahoo! Finance stock prices Application to official Italian COVID'19 data First Exploratory Data Analysis tasks on the Titanic dataset	Fri 27/05	17:00-19:00
4	Data pre-processing Application to breast cancer diagnosis	Mon 30/05	17:00-19:00
5	Unsupervised ML Application to breast cancer diagnosis	Mon 06/06	17:00-19:00
6	Supervised ML Application to breast cancer diagnosis	Fri 10/06	17:00-19:00
7	KNIME a graphical language for complex data analysis	Mon 13/06	17:00-19:00