Titolo insegnamento: Data Science **Modulo di Bioinformatica e Data Science**			
Title: Data science			
CFU: 6	**SSD: ING-INF/03**	**a.a. 2020-2021**	
Corso di laurea magistrale in Biotecnologie Agro-Ambientali e Alimentari ***Master's degree in Agro-Environmental and Food Biotechnology***			
Articolazione del corso (in ore): 42	**Lezioni:** 18	**Laboratorio:** 24	**Seminari:** 0
Docente: Edoardo Pasolli	**Tel. 0812539397**	**Email: edoardo.pasolli@unina.it**	

Anno di corso: Primo
Semestre: Secondo
Insegnamenti propedeutici previsti dal regolamento in vigore: Nessuno

RISULTATI DI APPRENDIMENTO ATTESI

Conoscenza e capacità di comprensione

La figura del "data scientist" - sempre più richiesta sia in ambito di ricerca che aziendale - si occupa di dare un senso all'enorme quantità di dati attualmente disponibile in molteplici campi, incluso quello delle biotecnologie agroalimentari e ambientali. Il corso si prefigge di dare allo studente le conoscenze e capacità di base per l'analisi dei dati attraverso l'uso di approcci statistici, l'implementazione di strumenti informatici e lo sviluppo di tecniche di apprendimento automatico (machine learning).

Data science is a growing inter-disciplinary field that uses scientific methods and algorithms to extract knowledge from the large amount of data that is currently available in multiple applications including the agro-environmental and food biotechnology sectors. The course aims at introducing the basic concepts for data analysis based on statistical approaches, implementation of (bio)informatic tools, and development of machine learning technologies.

Conoscenza e capacità di comprensione applicate

Il corso ha lo scopo di fornire un'introduzione sia teorica che pratica al campo della "data science". Le lezioni teoriche saranno accompagnate da attività in laboratorio aventi lo scopo di verificare e sperimentare gli argomenti trattati mediante la scrittura di semplici programmi informatici in ambiente R sotto la supervisione del docente con particolare enfasi ad applicazioni nell'ambito delle biotecnologie agroalimentari e ambientali e alla bioinformatica.

The course aims at introducing data science from both methodological and practical point of views. Theoretical lessons will be integrated with laboratory activities in which students will learn how to write code in R programming language. Emphasis will be given to examples in the agro-environmental, food biotechnology, and bioinformatics fields.

Eventuali ulteriori risultati di apprendimento attesi, relativamente a:

Lo studente svilupperà capacità di base di analisi dei dati che gli consentiranno di risolvere problemi di

diversa complessità e di gestire osservazioni numeriche, anche attraverso l'utilizzo di strumenti di comunicazione e rappresentazioni grafiche.

The student will acquire a basic knowledge in data analysis which will be beneficial to address problems of varying complexity and manage numerical observations. This will be done also by adopting communication tools and graphical representations.

PROGRAMMA

- Lezioni di teoria (3 CFU)
    - Introduzione alla "data science" per sistemi biologici; Acquisizione e pre-processamento del dato; Esplorazione e visualizzazione del dato (1 CFU)
    - Algoritmi di apprendimento automatico (1 CFU)
    - Hidden Markov model; L'era dei "big data"; Quiz e discussione finale (1CFU)
- Lezioni di laboratorio (3 CFU)
    - Introduzione a R & I primi comandi; Come costruire un Markdown & Lettura e scrittura del dato; Subsetting degli oggetti & Pacchetto dplyr (1 CFU)
    - Strutture di controllo & Pacchetto ggplot2; Analisi delle componenti principali & Multidimensional scaling (1 CFU)
    - K-means clustering & Hierarchical clustering; Pacchetto caret & Classificazione supervisionata; Regressione lineare e non lineare (1 CFU)

CONTENTS

- Theory classes (3 CFU)
    - Introduction to data science for biological systems; Data acquisition and pre-processing; Data exploration and visualization (1 CFU)
    - Machine learning algorithms (1 CFU)
    - Hidde Markow model; The big data era; Final quiz and discussion (1 CFU)
- Practice classes (3 CFU)
    - Introduction to R & The first commands; How to build a Markdown & Data reading and writing; Subsetting & Package dplyr (1 CFU)
    - Control structures & Package ggplot2; Principal component analysis & Multidimensional scaling (1 CFU)
    - K-means clustering & Hierarchical clustering; Package caret & Supervised classification; Linear and non-linear regression (1 CFU)

MATERIALE DIDATTICO

- Slide del corso
- "R programming for data science", Roger Peng, 2016
- "Exploratory data analysis with R", Roger Peng, 2016

FINALITA' E MODALITA' PER LA VERIFICA DI APPRENDIMENTO

L'esame finale mira a verificare i seguenti risultati di apprendimento:

- Conoscenza del processo complessivo alla base dell'analisi dei dati (acquisizione, pre-

processamento, modellazione/apprendimento automatico, comunicazione e visualizzazione dei dati e risultati)

- Rappresentazione dei dati in un contesto di “data science” (ad esempio campioni vs features)
- Tecniche di apprendimento automatico (ad esempio selezione delle feature, approcci supervisionati vs non-supervisionati, classificazione vs regressione)
- Conoscenza di base dell’ambiente R, implementazione di programmi per l’analisi dei dati, ed interpretazione dei risultati in ambito di biotecnologie agro-ambientali e alimentari

PURPOSES AND MODALITIES OF LEARNING VERIFICATION

- Knowledge of the overall process for data analysis (data acquisition, pre-processing, modelling and machine learning, communication and visualization of data and results)
- Data representation in a data science context (e.g., samples vs features)
- Machine learning methodologies (e.g., feature selection, supervised vs unsupervised approaches, classification vs regression)
- Basic knowledge of the R environment, implementation of code for data analysis, interpretation of results for agro-environmental and food biotechnology applications

Modalità di esame:

L'esame si articola in prova	Scritta e orale		Solo scritta	X	Solo orale	
Discussione di elaborato						
Altro, specificare						

In caso di prova scritta i quesiti sono (*)	A risposta multipla	4	A risposta libera	22	Esercizi numerici	

(*) E' possibile rispondere a più opzioni

NOTE DEL DOCENTE

Nessuna