

SummeR School-DATA MINING AND DATA ANALYSIS WITH R

PROPONENTE

Dott. Davide Gentilini Prof.ssa Luisa Bernardinelli

OBIETTIVI FORMATIVI

La gestione e l'analisi dei dati rappresenta uno dei fattori più importanti e simultaneamente più critici in molti ambiti lavorativi; spesso ci si avvale di strumenti e software proprietari che sono soventemente costosi o risultano essere in definitiva limitati nella loro applicazione.

Il corso ha l'obiettivo di fornire ai partecipanti le conoscenze necessarie e sufficienti per introdurre ed utilizzare il linguaggio e le potenzialità di R nel proprio lavoro.

R può essere definito come un sistema di analisi statistica e contemporaneamente un linguaggio ed un software. E' uno strumento open source potentissimo e largamente utilizzato, per l'analisi statistica dei dati, inoltre, essendo un vero e proprio linguaggio di programmazione, racchiude in se la potenzialità per creare e sviluppare in autonomia svariate applicazioni utili alla manipolazione, gestione ed analisi di ogni tipo di dato. Le sue caratteristiche principali comprendono infatti la semplicità nella gestione e manipolazione dei dati, la disponibilità di una suite di strumenti per calcoli su vettori, matrici ed altre operazioni complesse, l'accesso ad un vasto insieme di strumenti integrati e funzioni sviluppate da altri e resi disponibili per l'analisi statistica, la produzione di numerose potenzialità grafiche particolarmente flessibili, possibilità di adoperare un vero e proprio linguaggio di programmazione orientato ad oggetti che consente l'uso di strutture condizionali e cicliche, nonché di funzioni create dall'utente.

Offrendo un'elaborata introduzione alla programmazione con R questo corso ha lo scopo di intercettare le esigenze dei partecipanti focalizzandosi in particolar modo su alcuni aspetti fondamentali come la manipolazione e gestione dei dati, la loro analisi attraverso l'identificazione del test statistico più appropriato e la visualizzazione di dati e risultati utilizzando le potenzialità grafiche messe a disposizione da R.

Il corso prevede l'impiego di numerosi "dataset" ed esempi che possano essere familiari alle varie aree di interesse in modo da agevolare i partecipanti nella comprensione ed applicazione delle nozioni acquisite.

NUMERO DI ORE(CFU)/LEZIONI

Il corso avrà una durata di almeno 20 ore (5 CFU), e sarà suddiviso in 5 incontri.

Possono essere frequentati anche singoli incontri.

PERIODO DI SVOLGIMENTO

Il periodo di svolgimento del corso sarà dal 3 al 7 Giugno 2019 (summeR school) durante i mesi di giugno o luglio.

DOCENTI

Dott. Davide Gentilini, Dott. Antonino Oliveri

COMITATO SCIENTIFICO

Prof Luisa Bernardinelli, Dott. Antonino Oliveri, Dr Davide Gentilini, Prof.ssa Elisa Fazio

COMITATO ORGANIZZATORE

Dipartimento di Scienze del Sistema Nervoso e del Comportamento e Residenza Biomedica – Fondazione Collegio Universitario S. Caterina da Siena, Pavia

SEDE DEL CORSO

Residenza Biomedica –Fondazione Collegio Universitario S. Caterina da Siena, Pavia

POTENZIALI DOTTORATI INTERESSATI

Il corso ha l'intento di essere trasversale e di fornire competenze utili in ogni ambito che abbia a che fare con dati e con la necessità di elaborarli e gestirli. Per tale ragione il corso ha l'obiettivo di essere utile ad ogni tipologia di dottorato, in particolare macroarea di Scienze della vita, Bioingegneria e Bionformatica, Economics, Political Studies, DREAMT. Nell'attività pratica verranno utilizzati esempi e dati nei vari ambiti applicativi di interesse per i diversi Dottorati.

MODALITA' DI VERIFICA DELL'APPRENDIMENTO

Il grado di apprendimento verrà testato al termine di ogni lezione sottoponendo i partecipanti ad un test.

Il test sarà composto da una serie di 10 domande a scelta multipla riguardanti la parte teorica a cui verrà aggiunto un esercizio inerente le argomentazioni trattate.

PROGRAMMA

3 Giugno 2019 , 9.30-13.30

Titolo: “L’ambiente di programmazione R generalità e i principali oggetti”

Obiettivo della lezione è introdurre i partecipanti all’ambiente R, illustrandone le principali funzionalità, l’architettura e presentando i principali oggetti.

- Introduzione del Corso e Generalità
- Installazione di R e configurazione
- Nozioni preliminari sulla sua architettura sui pacchetti e sulle funzioni
- Utilizzo dei I pacchetti
- Utilizzo dell’Help
- Gli oggetti principali
 - I vettori
 - o Assegnazione Vettori (Vettori aritmetici, Logici, di Caratteri)
 - o Operazioni e funzioni per lavorare con i vettori
 - Matrici ed Array
 - o Operatori e funzioni per il calcolo Matriciale
 - Liste e Data Frames
 - o Operatori e funzioni per lavorare con Data Frames e Liste
- Esempi
- Esercitazione Pratica

4 Giugno 2019 , 9.30-13.30

Titolo: “Data import e data mining con R”

Obiettivo della lezione è insegnare ai partecipanti ad importare dati di varia tipologia e formato nell’ambiente R. La lezione ha inoltre lo scopo di insegnare “data mining” ovvero ad utilizzare R per visualizzare gestire, estrarre e manipolare i dati. La lezione tratterà le funzioni grafiche di R in modo che i partecipanti al corso acquisiscano le conoscenze base per poter visualizzare i propri dati e i propri risultati in modo autonomo. Esempi ed esercizi guidati avranno lo scopo di verificare e consolidare le nozioni.

- Importazione e pulizia di data-set
 - Funzioni di importazione dei dati in base alla loro natura e formato
 - Funzioni di visualizzazione e manipolazione dei dati
- Quality Control dei dati
- Trattamento dei dati mancanti
- Imputazione
- Metodi di riclassificazione delle variabili
- Funzioni per la manipolazione dei dati
- L’ambiente grafico
 - Pacchetto Base
 - Pacchetto Grid
 - Pacchetto ggplot2
 - I parametri grafici
 - o Funzioni di alto e basso livello
 - o Funzioni interattive
 - Tipi di grafici e funzioni
 - o Grafici standard
 - o Grafici speciali
 - o Grafici dinamici
- Classificazione del tipo di dati e discussione sulla modalità di visualizzazione
- Esempi
- Esercitazione Pratica

5 Giugno 2019 , 9.30-13.30

Titolo: “Statistica di base con R”

Obiettivo della lezione è entrare nel merito della statistica descrittiva/inferenziale e fornire gli strumenti essenziali per poter scegliere e applicare il metodo statistico più appropriato

- Misure descrittive univariate/bivariate
- Test statistici
 - Test parametrici su media, differenza fra medie, proporzione, differenza fra proporzioni e correlazione
 - Test non parametrici su media, differenza fra medie, differenza fra mediane, proporzione, differenza fra proporzioni e correlazione
 - Test di associazione per tabelle 2x2 e JxK
- Esempi
- Esercitazione Pratica

6 Giugno 2019 , 9.30-13.30

Titolo: “Statistica di base con R”

Obiettivo della lezione è entrare nel merito della statistica descrittiva/inferenziale e fornire gli strumenti essenziali per poter scegliere e applicare il metodo statistico più appropriato

- Analisi della varianza ad una via, a due vie e problema dei confronti multipli
- Regressione lineare semplice e multipla
- Regressione logistica
- Analisi delle componenti principali
- Analisi dei cluster

- Esempi
- Esercitazione Pratica

7 Giugno 2019 , 9.30-13.30

Titolo: “Principi Base di programmazione con R”

Obiettivo della lezione: esplorare i rudimenti di programmazione “Object Oriented” con R e fornire le basi affinché il partecipante sia in grado di sviluppare semplici programmi che possano automatizzare alcuni processi di lavoro o di analisi. La lezione offre anche una visione sulla possibilità di sviluppare applicazioni web e on site.

- Istruzioni condizionali
- Istruzioni iterative
- Definizione e richiamo di funzioni
- Applicazioni in R e Shiny

- Esempi
- Esercitazione Pratica