TerreLogiche®

www.terrelogiche.com

Corso “Statistica multivariata e machine learning con R”

Introduzione ai principali strumenti di analisi, classificazione e processamento di dataset complessi con utilizzo di tecniche di statistica multivariata e machine learning

>> 15% di sconto per chi si iscrive con almeno 30 giorni di anticipo
>>10% di sconto per iscritti a Ordini e Associazioni professionali, Categorie Educational e dipendenti di Pubbliche Amministrazioni

---

Informazioni ed iscrizioni: www.terrelogiche.com

TerreLogiche®

"**Statistica multivariata e machine learning con R**" di TerreLogiche è un corso di formazione **online (live streaming)** introduttivo ai principali strumenti per l'analisi di dataset complessi e alle tecniche di apprendimento automatico in ambiente R.
In particolare, all'interno del modulo saranno affrontate le più comuni tecniche di analisi multivariata (***Principal Components Analysis***, ***Canonical Correspondence Analysis***, ***Multidimensional Scaling***, ***Positive Matrix Factorization***, ***Cluster Analysis***) e di machine learning (***RandomForest***, ***Gradient Boosting Machines***).

La **statistica multivariata** si occupa di **estrarre informazioni e costruire modelli predittivi** in situazioni in cui la mole e multidimensionalità dei dati da analizzare e la tipologia di informazioni sono tali da non poter essere gestiti efficacemente con gli strumenti della statistica classica.
Il termine ***machine learning*** comprende differenti metodologie, sviluppate negli ultimi decenni, che utilizzano criteri statistici per migliorare la performance di un **algoritmo di apprendimento automatico** nell'identificare e descrivere pattern nei dati oggetto di analisi. Gli algoritmi di apprendimento automatico sono utilizzati in numerosi settori tra loro molto differenti e, in generale, in qualsiasi contesto dove gli algoritmi convenzionali non possono svolgere i compiti richiesti in modo efficiente.

In un mondo dove la disponibilità di dati sta crescendo sempre più rapidamente è fondamentale avere la possibilità di utilizzare strumenti e tecnologie che consentano in maniera efficace di svolgere analisi critiche e informative generando, contestualmente, modelli predittivi un tempo impossibili da realizzare.
Le tecniche illustrate all'interno del corso sono ampiamente utilizzate per **risolvere importanti problematiche in settori di ricerca molto differenziati** (es. biologia, epidemiologia, scienze ambientali, marketing, riconoscimento vocale, filtraggio informazioni, ecc.) e, in generale, in tutti quegli ambiti dove la quantità di dati disponibili è aumentata notevolmente negli ultimi decenni con la necessità di conoscenze specifiche per riuscire a estrarre informazioni da dati disomogenei ed eterogenei accompagnati da un elevato livello di "rumore".
Lo scopo principale di queste tecniche è infatti quello di visualizzare o estrarre le informazioni desiderate da dati multivariati, in presenza di "rumore", **attraverso l'analisi di dataset complessi**, la **riduzione della dimensionalità**, la **classificazione** e la **predizione di risposte**.

Il modulo formativo fornisce, attraverso **esercitazioni pratiche**, le **conoscenze operative necessarie per la gestione dei dati multivariati e il loro processamento utilizzando il software Open Source R**, uno dei più potenti e flessibili sistemi attualmente disponibili per l'elaborazione statistica dei dati e la loro rappresentazione grafica.

Contenuti e obiettivi del corso

Il modulo formativo permetterà ai partecipanti di acquisire le **nozioni fondamentali per lo studio di dataset complessi e l'applicazione di alcune tecniche di analisi multivariata attraverso l'utilizzo dell'ambiente R**. Saranno inoltre illustrate **potenzialità e procedure di ottimizzazione di due moderni algoritmi di machine learning**, fornendo ai partecipanti una solida base di istruzioni operative per l'utilizzo di tali strumenti.
L'approccio metodologico del corso e l'organizzazione dei contenuti sono basati su un **flusso di lavoro ben collaudato con la possibilità, per i partecipanti, di ripetere in ogni momento le operazioni eseguite dal docente e lo svolgimento di numerose esercitazioni pratiche riguardanti le tecniche illustrate**. Le tecnologie utilizzate dalla sessione formativa saranno totalmente **Open Source**, gratuite e liberamente scaricabili.

Nella prima parte del corso, dopo un breve richiamo di alcune nozioni statistiche di base, saranno affrontate le più comuni **tecniche di analisi multivariata** e in particolare:

- **Principal Components Analysis (PCA)**: una tecnica multivariata che ha come obiettivo principale quello di ridurre la dimensionalità di un dataset multivariato tenendo conto della maggior parte possibile della variazione originale. È utilizzata in moltissimi ambiti, dalla finanza alla ricerca medica, geologica e chimica soprattutto per rappresentare i set di dati tramite un numero inferiore di nuove variabili tra loro non correlate.

- **Canonical Correlation Analysis (CCA)**: una tecnica che consente di dedurre informazioni da matrici di covarianza incrociata. Un uso tipico della correlazione canonica, nel contesto sperimentale, è quello di analizzare due insiemi di variabili e verificare gli elementi in comune tra i due insiemi stessi (es. chiarire le relazioni tra comunità biologiche e il loro ambiente).
- **Non-metric MultiDimensional Scaling (NMDS)**: tecniche multivariate con l'obiettivo principale di ridurre la dimensionalità di un set di dati multivariato utilizzando matrici di prossimità tra le osservazioni. Trova utilizzi nelle scienze ambientali, in sociologia, in psicologia e nel marketing quando il focus delle analisi è lo studio delle distanze/differenze descritte attraverso molteplici variabili.
- **Positive Matrix Factorization (PMF)**: una tecnica di analisi fattoriale multivariata utilizzata per la scomposizione di problemi complessi nelle sue componenti positive. È applicata con successo, tra gli altri, dalla US Environmental Protection Agency per la valutazione di dataset ambientali. Può essere utilizzata anche per la classificazione di immagini, la scomposizione dei consumi elettrici nelle sue componenti oppure l'identificazione della "ricetta" di un colore creato dalla mescolanza di altri.
- **Cluster Analysis**: un'ampia gamma di metodi numerici con l'obiettivo comune di definire o scoprire gruppi di osservazioni tra loro omogenee separandole da altri gruppi appartenenti al campione oggetto di studio. Nel marketing viene utilizzata per suddividere la popolazione generale dei consumatori in segmenti di mercato ma può anche essere applicata per classificare immagini, effettuare analisi di reti sociali e per l'identificazione di anomalie in moltissimi altri settori.

Successivamente sarà illustrato l'utilizzo di due fondamentali algoritmi di machine learning:

- **RandomForest**: un algoritmo comunemente utilizzato che combina l'output di più strutture ad albero decisionali per raggiungere un unico risultato. La facilità d'uso e la flessibilità ne hanno favorito la diffusione, in quanto può gestire sia i problemi di classificazione che quelli di regressione. Data la sua flessibilità è utilizzato in svariati campi: es. valutazione della predisposizione a diverse malattie, identificazione di segmenti di consumatori, valutazione delle relazioni tra ambiente e organismi viventi, ecc.
- **Gradient Boosting Machines (GBMs)**: una famiglia di tecniche di apprendimento automatico che hanno dimostrato successo in un'ampia gamma di applicazioni. Sono altamente personalizzabili in diversi aspetti e la procedura di apprendimento adatta in modo consecutivo nuovi modelli per fornire una stima più accurata della variabile di risposta.

Per ogni tipologia di analisi illustrata saranno definite le caratteristiche e le principali potenzialità, presentata la definizione in ambiente R, l'ottimizzazione, l'interpretazione degli output e la sintesi numerica e grafica degli stessi.

Cos'è R

R è un **ambiente Open Source per l'analisi statistica dei dati e la loro rappresentazione grafica**. Può essere installato su **piattaforme MS Windows, Linux** e **MacOS**. Il programma offre una vasta gamma di tecniche di analisi statistica e grafica (es. modellizzazione lineare e non lineare, test statistici classici, analisi delle serie temporali, la classificazione, il clustering, ecc..). R è un sistema integrato di utilities per la manipolazione dei dati, il calcolo e la visualizzazione grafica. Include:

- Una gestione efficace dei dati;
- Un insieme di operatori per i calcoli su array, in particolare matrici, ed una raccolta di strumenti intermedi per l'analisi dei dati;
- Strutture grafiche per l'analisi dei dati e la loro visualizzazione sia su schermo che su supporto cartaceo;
- Un linguaggio di programmazione orientato ad oggetti semplice e ben sviluppato che comprende, ad esempio, istruzioni condizionali, loop e funzioni ricorsive.

A chi è rivolto questo corso

Il corso è rivolto a professionisti, ricercatori, tecnici di Pubbliche Amministrazioni, studenti universitari, insegnanti e in genere a tutti coloro che intendono ampliare le loro conoscenze nell'ambito dell'elaborazione statistica dei dati.

Livello e requisiti di accesso

Per la partecipazione al corso sono richieste le seguenti conoscenze:

- Buona conoscenza del proprio sistema operativo (MS Windows, Linux, MacOS) e della relativa gestione di file e cartelle;
- Conoscenza di nozioni di base di statistica;
- Conoscenza di base del linguaggio R (in particolare, essere in grado di creare e manipolare oggetti di diverso tipo, importare ed esportare informazioni). Nonostante all'interno del corso sia previsto un breve modulo introduttivo all'uso di tecniche statistiche in R, le predette conoscenze di base sono richieste. In assenza di esse consigliamo la partecipazione propedeutica al corso "**Statistica con R (Base**)"

Tipologia e modalità del corso

Corso interattivo con lezioni frontali **in aula** o **online in modalità live streaming**.
Al momento questo corso viene erogato esclusivamente in modalità online (live streaming). Ricreiamo nelle aule virtuali l'esperienza formativa proposta nei corsi in presenza quindi **approccio pratico alle tematiche affrontate**, **esercitazioni e laboratorio assistito** con una **forte interazione tra docente e discente** e ampio spazio ai quesiti dei partecipanti.

Personale docente

Il corso è tenuto da docente senior con larga esperienza nel trattamento statistico univariato e multivariato dei dati in ambiente R.

Dotazione informatica

È necessario l'utilizzo di notebook personale e di connessione Internet stabile e di adeguata velocità. Non sono richiesti particolari requisiti hardware (RAM ameno 4 GB).

Sede del corso

Questo corso è attualmente erogato in modalità online (live streaming). Consulta il calendario su www.terrelogiche.com.

Durata

20 ore
Per il dettaglio degli orari di svolgimento, consultare le specifiche della singola sessione: https://www.terrelogiche.com/formazione-terrelogiche/scopri-i-corsi/statistica-multivariata-e-machine-learning-con-r.html

Crediti Formativi

Consultare le specifiche della singola sessione per ulteriori informazioni.

Costi e riduzioni

Consulta il **calendario dei corsi** con i relativi costi su www.terrelogiche.com.
Tutti coloro che si iscriveranno al corso con almeno 30 giorni di anticipo rispetto alla data della sessione formativa avranno diritto ad uno **sconto del 15%** sul prezzo di listino. È inoltre previsto uno **sconto del 10%** sul prezzo di listino per gli iscritti a Ordini ed Associazioni professionali (Legge 4 del 14 gennaio 2013), Categorie Educational e dipendenti di Pubbliche Amministrazioni. Gli sconti non sono cumulabili se non diversamente concordato. Consultare i dettagli nella sezione **Agevolazioni** del sito www.terrelogiche.com.

Agevolazioni fiscali

I costi della formazione sono **interamente deducibili (100%) per aziende e professionisti** (art. 54 c. 5 TUIR DPR 917/1986). Solamente per questi ultimi è fissato un tetto annuo di euro 10.000 (comprensivo di spese di soggiorno e trasferta), per le aziende non esistono limiti annui. L'IVA è 100% detraibile. Inoltre, le Pubbliche Amministrazioni hanno diritto all'esenzione IVA riferita ad attività formative (DPR 633/72).

Modalità di iscrizione

La procedura di iscrizione è molto semplice. Le istruzioni sono indicate nella Sezione Formazione su www.terrelogiche.com.

Attestati di partecipazione e profitto

Al termine della sessione formativa verrà rilasciato a tutti i partecipanti che hanno **frequentato almeno il 70%** del monte ore totale un **attestato di partecipazione** numerato e personale con specificate il numero di ore del corso e le principali tematiche affrontate.

È inoltre previsto lo svolgimento (opzionale) di un **test finale di valutazione dell'apprendimento** con domande a risposta multipla, che si intende **superato** fornendo **almeno l'80% delle risposte corrette**. Il superamento del test sarà certificato su un **attestato di partecipazione e profitto**, documento utile per **arricchire il proprio curriculum** in quanto documenta che sono state acquisite le competenze e le conoscenze previste dal corso frequentato.

Il test finale di valutazione **non è obbligatorio** e **non comporta un aumento del costo di iscrizione**.

Vantaggi del corso e materiale fornito

- Formazione erogata secondo gli **standard di qualità ISO 9001:2015**;
- Aule (virtuali) con **numero limitato di posti** per una migliore efficacia didattica;
- **Ampio materiale didattico in formato digitale scaricabile dal cloud TerreLogiche** (slides, dataset, documentazione e manualistica riguardante i software e le tematiche affrontate);
- **Attestato di partecipazione** numerato e personale con specificate il numero di ore del corso e le principali competenze acquisite, rilasciato ai partecipanti che hanno **frequentato almeno il 70%** del monte ore totale. **Attestato di partecipazione e profitto**, rilasciato a seguito del **superamento del test finale** di valutazione dell'apprendimento. Su richiesta l'attestato viene erogato anche in lingua inglese;
- **Supporto tecnico** per eventuali problematiche di installazione e configurazione dei software utilizzati;
- **Test di connessione**: nei giorni precedenti il corso sarà effettuato un breve test di connessione con il docente (opzionale), per illustrare le funzionalità della piattaforma utilizzata, verificare la velocità delle connessioni e risolvere eventuali problemi tecnici dei partecipanti nella configurazione e installazione dei software;
- Buoni sconto di TerreLogiche.

Programma del corso

- **Breve riepilogo delle procedure di base in ambiente R**
  - Tipologie di variabili
  - Esplorazione dei dati
  - Teoria dei test
  - Test statistici parametrici
  - Test statistici non parametrici
  - Interpretazione degli output
  - Assunti della regressione
  - Processi di selezione dei modelli

- Procedure di test delle ipotesi
- Metodi grafici di validazione dei modelli di regressione

- **Introduzione alla statistica multivariata**
  - Gestione dei dati mancanti
  - Matrici di covarianza e correlazione
  - Calcolo delle distanze multivariate
  - Metodi grafici per la visualizzazione di dataset multivariati
  - Plots tridimensionali

- **Principal Components Analyses (PCA)**
  - Introduzione al problema della riduzione della dimensionalità
  - Quando scalare le variabili prima dell'analisi
  - Matrici di covarianza e correlazione
  - Identificazione del numero di componenti rilevanti
  - Esempi pratici di utilizzo della PCA
  - Estrazione delle informazioni dalla PCA
  - Visualizzazione dei risultati di PCA
  - Utilizzare le componenti principali come variabili indipendenti
  - Canonical Correlation Analysis (CCA)

- **Non-metric MultiDimensional Scaling (NMDS)**
  - Un approccio differente per ridurre la dimensionalità
  - La rappresentazione spaziale di matrici di distanza
  - Descrizione delle principali misure di distanza multivariata
  - Esempi pratici di utilizzo della NMDS
  - Estrazione delle informazioni dalla NMDS
  - Test per il confronto di posizione e dispersione dei gruppi
  - Visualizzazione dei risultati di NMDS

- **Positive Matrix Factorization (PMF)**
  - Riduzione della dimensionalità, un approccio differente
  - Introduzione al pacchetto "NMF"
  - Inizializzazione del modello
  - Ricerca del numero e della composizione ottimale delle sorgenti
  - Composizione delle sorgenti e contributo delle sorgentiù
  - Interpretazione degli output

- **Cluster Analysis**
  - Ordinare cose simili in categorie
  - Individuazione del numero di gruppi ottimale
  - Agglomerative hierarchical techniques
  - k-means clustering
  - Model-based clustering (finite mixture densities or latent class cluster analysis)
  - Introduzione al pacchetto "*mclust*"
  - Introduzione al pacchetto "*ClusterR*"

- **Introduzione alle tecniche di apprendimento automatico**
  - Teoria dell'apprendimento
  - Classificazione e regressione
  - Tipologie di approcci

- **Random Forest (RF)**
  - Introduzione al pacchetto "*randomForest*"
  - Esempio pratico e valutazione dei risultati
  - Quantile Random Forest, introduzione al pacchetto "*quantregForest*"

- **Gradient Boosting Machines (GBMs)**
  - Introduzione alle tecniche Gradient Boosting Machines
  - Vantaggi e svantaggi della tecnica GBMs
  - Definizione dei parametri del modello
  - Ottimizzazione dei parametri del modello
  - Introduzione al pacchetto "*gbm*"
  - Esempio pratico con l'utilizzo del pacchetto "*gbm*"
  - Introduzione al pacchetto "*xgboost*" e "*vtreat*"
  - Esempio pratico con l'utilizzo del pacchetto "*xgboost*"

Feedback

I corsi di TerreLogiche sono da molti considerati i migliori in Italia per qualità erogata, costi accessibili e per il forte approccio applicativo decisamente adeguato alla realtà lavorativa. I nostri sondaggi effettuati immediatamente dopo il corso e a campione a distanza di alcuni mesi hanno rivelato un'altissima percentuale di gradimento e soddisfazione. I **feedback** sui corsi di TerreLogiche sono al **99,8% positivi dal 1998**.