

# Fondamenti di Analisi Dati e Laboratorio

## Prima Prova in Itinere: Analisi Dati Esplorativa e Inferenziale

### 1. Obiettivo della prova

L'obiettivo di questo assignment è applicare le metodologie di analisi dati acquisite durante il corso per condurre un'analisi completa di un dataset. Partendo da dati grezzi, ogni studente singolo o gruppo dovrà esplorare, pulire, visualizzare e, infine, interrogare i propri dati attraverso l'uso di tecniche statistiche.

Lo scopo finale è trasformare i dati in conoscenza: identificare pattern, scoprire relazioni significative tra le variabili e validare ipotesi utilizzando un approccio statistico. L'analisi non deve essere una semplice esecuzione di comandi, ma una **narrazione supportata da evidenze quantitative e grafiche**. Ad ogni dataset assegnato saranno fornite una serie di domande alla quale sarà obbligatorio rispondere.

### 2. Fasi dell'Analisi Dati

Ogni studente dovrà strutturare il proprio elaborato (un Jupyter Notebook) seguendo un prototipo di analisi dati che includa le seguenti fasi fondamentali.

#### Fase 2.1: Acquisizione e Comprensione del Dataset (Data Understanding)

Il primo passo consiste nel caricamento del dataset assegnato e in una sua prima ispezione. È fondamentale comprendere il contesto dei dati.

- **Caricamento:** Importare il dataset in una struttura dati.
- **Ispezione Iniziale:** Verificare le dimensioni (numero di righe e colonne), i nomi delle colonne e i tipi di dato (Dtypes) inferiti automaticamente.
- **Dizionario dei Dati:** Comprendere a fondo il significato di ogni variabile (attributo). Cosa rappresenta? Qual è la sua unità di misura? È una variabile categorica (nominale, ordinale) o numerica (discreta, continua)?
- **Definizione degli obiettivi dell'analisi dei dati:** formulare 4-5 domande sulla base delle quali strutturare l'analisi dei dati. Le domande devono essere relative al fenomeno rappresentato nei dati e possono essere rifinite nelle successive fasi del progetto. Ad esempio, in un dataset clinico sul diabete, l'analisi può essere guidata da domande come:
  - Quali sono i principali fattori che influenzano la diagnosi di diabete?

- Esiste una relazione tra età e parametri metabolici come glucosio e BMI?
- Il numero di gravidanze influisce sui livelli glicemici o sulla diagnosi?
- Il BMI è associato a un rischio maggiore di diabete?

## Fase 2.2: Pulizia e Preprocessing (Data Cleaning & Preparation)

Questa fase è cruciale per garantire l'affidabilità delle analisi successive. Considerare ad esempio:

- **Gestione Valori Mancanti (NaN):**
  - Identificare le variabili con dati mancanti e quantificare la loro incidenza (es. percentuale di NaN per colonna).
  - Scegliere una strategia e motivarla:
- **Gestione Duplicati:** Identificare ed eliminare eventuali righe completamente duplicate.
- **Correzione Inconsistenze e Formattazione:**
  - Verificare che i tipi di dato siano corretti (es. convertire colonne di prezzo da stringhe a numeri, colonne di date da stringhe a oggetti datetime).
  - Correggere eventuali errori di battitura o formati incoerenti nelle variabili categoriche (es. "USA", "U.S.A.", "Stati Uniti" dovrebbero essere uniformati).
- **Gestione Outlier (Valori Anomali):**
  - Identificare visivamente gli outlier (es. tramite box plot).
  - Analizzare se si tratta di errori di inserimento (da correggere o rimuovere) o di valori estremi ma legittimi (da mantenere o trasformare, es. con una trasformazione logaritmica). Motivare la scelta.

## Fase 2.3: Analisi Esplorativa

L'obiettivo di questa fase è esplorare i dati e analizzarne le caratteristiche principali. A seconda dei dati, considerare le seguenti tecniche:

- **Analisi Univariata (analisi di una variabile alla volta):**
  - **Per variabili Numeriche (Continue/Discrete):**
    - **Indici Statistici:** Calcolare media, mediana, moda (per capire la tendenza centrale), deviazione standard, varianza, range e range interquartile (IQR) (per capire la dispersione e variabilità).
    - **Visualizzazioni:** Usare **Istogrammi** (per visualizzare la forma della distribuzione) e **Box Plot** (per identificare quartili, mediana e outlier).
  - **Per variabili Categoriche (Nominali/Ordinali):**
    - **Quantitative:** Creare tabelle di frequenza (assolute e relative) per capire la distribuzione delle categorie.
    - **Visualizzazioni:** Usare **Grafici** opportuni per mostrare le frequenze.
- **Analisi Multivariata (analisi delle relazioni tra più variabili):**
  - **Relazione Numerica vs. Numerica:**

- **Visualizzazione:** Usare **Plot adeguati** per identificare pattern (lineari, non lineari, assenza di relazione).
- **Quantitative:** Calcolare **Correlazioni** tra i dati.
- **Relazione Numerica vs. Categorica:**
  - **Visualizzazione:** Usare **Plot** per confrontare le distribuzioni della variabile numerica tra i diversi gruppi.
  - **Quantitative:** Calcolare gli indici statistici (es. media, mediana) della variabile numerica *raggruppati* per la variabile categorica, correlazioni.
- **Relazione Categorica vs. Categorica:**
  - **Quantitative:** Creare **Tabelle di Contingenza** (Tabelle a doppia entrata) per mostrare le frequenze incrociate. Calcolare statistiche Chi-Quadrato e V di Cramers.
  - **Visualizzazione:** Usare **Grafici a Barre Impilati** (Stacked Bar Charts) o **Raggruppati** (Grouped Bar Charts) per confrontare visivamente le proporzioni.

## Fase 2.4: Inferenza Statistica

Dopo aver esplorato i dati e identificato pattern o differenze interessanti (es. "sembra che il Gruppo A abbia una media più alta del Gruppo B"), questa fase serve a determinare se tali osservazioni sono *statisticamente significative* o se potrebbero essere dovute semplicemente al caso.

- **1. Formulare una Domanda:** Partendo da un'osservazione emersa durante l'analisi esplorativa (EDA), formulare una domanda chiara. (*Es. "La differenza di prezzo medio che ho osservato tra i prodotti 'bio' e 'standard' è reale o è solo una fluttuazione casuale del mio campione?"*).
- **2. Intervalli di confidenza:** se opportuno, usare stime intervallari per misurare e confrontare quantità sul campione.
- **3. Scegliere un Test Adeguato:** In base alla domanda e al tipo di dati che si stanno confrontando, selezionare un test statistico appropriato per validare l'ipotesi.
  - *Esempio 1 (Numerica vs. Categorica a 2 livelli):* Per confrontare le medie di due gruppi (come nell'esempio 'bio' vs 'standard') si può usare un **t-test per campioni indipendenti**.
  - *Esempio 2 (Categorica vs. Categorica):* Per verificare se esiste un'associazione tra due variabili categoriche (es. 'Regione di provenienza' e 'Tipo di prodotto acquistato') si può usare un **Test Chi-Quadrato** (chi quadrato).
- **4. Eseguire il Test e Interpretare i Risultati:**
  - Eseguire il test statistico scelto utilizzando un'opportuna libreria.
  - Interpretare il risultato del test per determinare la significatività statistica della

- propria osservazione.
- **Conclusione Pratica:** Sulla base del risultato del test, trarre una conclusione chiara e contestualizzata al problema. (*Es. "Abbiamo trovato prove statisticamente significative per affermare che esiste una reale differenza di prezzo tra i prodotti 'bio' e 'standard' in questo campione" oppure "Non abbiamo trovato prove sufficienti per concludere che la differenza di prezzo osservata sia statisticamente significativa, potrebbe essere dovuta al caso"*).

### 3. Consegna

La consegna consiste in un unico **Jupyter Notebook** (file .ipynb) che va consegnato via mail al docente entro la data della successiva prova in itinere. Questo notebook deve essere un documento auto-esplicativo che combini codice eseguibile e testo descrittivo.

- **Codice:** Tutto il codice Python utilizzato per le fasi di pulizia, analisi e inferenza deve essere presente e commentato.
- **Visualizzazioni:** Tutti i grafici devono essere leggibili, dotati di titoli appropriati, etichette sugli assi e, se necessario, una legenda.
- **Celle Markdown:** Il notebook deve includere un report testuale (scritto in celle Markdown) che guida il lettore attraverso l'analisi e gli snippets di codice presenti.  
Questo report deve:
  - Motivare le scelte di pulizia e preprocessing.
  - Descrivere e commentare i risultati dell'analisi esplorativa (cosa si evince dai grafici e dagli indici statistici?).
  - Dettagliare la formulazione, l'esecuzione e l'interpretazione dei test di inferenza statistica.