

# Fondamenti di Analisi Dati e Laboratorio

## Seconda Prova in Itinere: Modellazione Statistica e Machine Learning

### 1. Obiettivo della prova

L'obiettivo di questo secondo assignment è estendere l'analisi iniziata nella prima prova, passando dalla fase esplorativa a quella di **modellazione**. Gli studenti dovranno utilizzare il dataset già assegnato per costruire modelli dei dati.

In questa fase, l'obiettivo si sdoppia:

1. **Spiegare (Approccio Statistico):** Utilizzare modelli di regressione per quantificare le relazioni tra le variabili, testare ipotesi e comprendere l'impatto dei fattori in gioco (es. "Di quanto aumenta la pressione sanguigna per ogni anno di età in più?").
2. **Predire (Approccio Machine Learning):** Costruire sistemi automatici in grado di stimare valori futuri o classificare nuove istanze con l'obiettivo di generare valore pratico (es. "Possiamo predire se un paziente ha il diabete solo guardando i valori del sangue, risparmiando esami più invasivi?").

Il lavoro dovrà essere svolto proseguendo il **Jupyter Notebook** della prima prova, aggiungendo nuove sezioni.

### 2. Fasi dell'Analisi: Modellazione e Predizione

Gli studenti dovranno aggiungere al notebook le seguenti parti, applicando le metodologie viste a lezione.

#### Analisi Statistica e Regressione (Inference)

In questa parte, l'attenzione è posta sulla significatività statistica e sull'interpretazione dei coefficienti. Non stiamo ancora cercando di fare la "miglior predizione possibile", ma di capire **come** le variabili si influenzano a vicenda.

#### Attività richieste:

- **Selezione delle Variabili:** Sulla base dell'analisi esplorativa precedente, individuare una o più variabili target (dipendenti) e un set di predittori (indipendenti) che abbiano senso logico e teorico.
- **Applicazione dei Modelli:** A seconda della natura della variabile target, applicare uno o più dei seguenti modelli utilizzando un approccio statistico (libreria `statsmodels`):
  - *Regressione Lineare:* Se la target è numerica continua.
  - *Regressione Logistica / Multinomiale:* Se la target è categorica (binaria o multiclasse).
- **Interpretazione e Diagnostica:**
  - Discutere i **coefficienti** ottenuti: che significato hanno nel mondo reale? (Es. "Un incremento unitario della variabile X è associato a un aumento di Y pari a beta, mantenendo costanti le altre variabili").

- Analizzare i **p-value**: quali predittori sono statisticamente significativi?
- Analizzare gli **intervalli di confidenza**.
- Valutare la bontà di adattamento (R-squared, Pseudo R-squared) e, se necessario, verificare le assunzioni del modello (es. analisi dei residui per la regressione lineare).

## Analisi Predittiva e Machine Learning

In questa parte cambiamo prospettiva: l'obiettivo è creare un **modello predittivo** che funzioni su dati mai visti.

### 4.1 Definizione del Problema e del Valore Pratico

Per identificare il problema predittivo da risolvere, porsi la domanda "**Che strumento automatico posso creare a partire da questi dati?**"

Il progetto dovrà contenere una discussione che miri a rispondere alla domanda di cui sopra:

- Descrivere uno scenario applicativo.
- Spiegare perché è utile risolvere questo problema.
- Quantificare (anche ipoteticamente) il valore aggiunto.
  - *Esempio 1*: Predire se un paziente ha il diabete basandosi su dati anagrafici e prelievi non invasivi. *Valore*: Risparmio di test costosi e diagnosi precoce.
  - *Esempio 2*: Stimare il prezzo di vendita di una casa. *Valore*: Supporto decisionale automatico per agenzie immobiliari.
- **Nota**: Se il dataset non ha una variabile target ovvia per la classificazione, è possibile crearla. *Esempio*: Discretizzare una variabile continua (es. "Pressione Sanguigna") in classi ("Bassa", "Normale", "Alta") per trasformare un problema di regressione in uno di classificazione, se questo approccio è ritenuto più opportuno (es. non ci importa predire l'esatto valore di pressione, ma capire se il paziente può soffrire di pressione alta).

### 4.2 Setup Sperimentale

- **Data Splitting**: Suddividere rigorosamente i dati in **Training Set**, **Validation Set** (opzionale, se si usa Cross-Validation, anche a seconda della dimensione del dataset) e **Test Set**.
- **Preprocessing per ML**: Applicare le trasformazioni necessarie (es. dummy variables per variabili categoriche, Scaling/Normalizzazione per algoritmi sensibili alle distanze).

### 4.3 Modellazione e Selezione

Applicare diversi algoritmi di Machine Learning visti a lezione (usando `scikit-learn`). Esempi di algoritmi sono:

- **Regressione**: Lineare, Polinomiale (per catturare non-linearità).
- **Classificazione**: Regressione Logistica, Softmax Regression, Naive Bayes, LDA (Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis), KNN (K-Nearest Neighbors).

**Requisiti:**

- Scegliere gli algoritmi più adatti al problema definito in 4.1. La scelta deve essere guidata dalla conoscenza maturata sui dati, in riferimento alle caratteristiche di ciascun algoritmo (es. evitare KNN per dati ad altissima dimensionalità).
- Eseguire una ricerca degli **iperparametri** (es. il grado del polinomio, il  $k$  in KNN, la penalizzazione in Lasso/Ridge) utilizzando tecniche come Grid Search o Cross-Validation sul Training Set.
- Motivare la scelta degli algoritmi.

#### 4.4 Valutazione e Confronto

- Calcolare le metriche di performance appropriate sul **Test Set**, ad esempio:
  - *Per Regressione*: MSE, MAE, RMSE, R2.
  - *Per Classificazione*: Accuracy, Precision, Recall, F1-Score, Matrice di Confusione, ROC curve.
- **Confronto Critico**: Confrontare i modelli tra loro.
  - *Esempio*: "La Regressione Polinomiale di grado 2 ha ridotto l'errore rispetto alla Lineare, ma il grado 10 ha mostrato chiaro overfitting."
  - *Nota*: È utile testare anche un modello "sbagliato" o semplice (baseline) per dimostrare quanto guadagno portano i modelli più appropriati (es. confrontare LDA con KNN).

### 3. Conclusioni Parziali e Consegna

Il notebook dovrà concludersi con una sezione di sintesi che riassume:

1. Le principali relazioni statistiche scoperte.
2. Le performance del miglior modello predittivo ottenuto e la sua applicabilità nello scenario ipotizzato.

#### Istruzioni Generali:

- **Non è obbligatorio applicare TUTTI gli algoritmi elencati**: Applicare quelli che hanno senso per i propri dati e per il problema formulato.
- **Narrazione**: Le celle di testo (Markdown) sono importanti quanto il codice. Spiegate il *perché* delle vostre scelte.