

## *Elenco di errori popolari degli studenti*

Segue un florilegio di errori che lo studente diligente avrà cura di evitare di commettere nella stesura del report.

### Errori di carattere generale

Scelto uno specifico data set, scopo dell'analisi statistica è quello di estrarre quanta più informazione utile alla soluzione di problemi. Pertanto, per ogni singolo data set, è raccomandabile applicare tutte quelle tecniche a disposizione, che possano fungere allo scopo. Gli errori comuni sono:

1. Omettere analisi statistiche rilevanti, non estraendo in tal modo tutta l'informazione possibile
2. Non ottenere tutte le possibili conclusioni scientifiche dall'output prodotto.

### Errori specifici

#### Regressione

1. Utilizzare regressori quasi collineari

2. Non valutare la significatività statistica dei coefficienti
3. Non mettere a confronto un modello generale con uno ristretto (con meno regressori), quindi non utilizzare il corrispondente test F
4. Non verificare se le ipotesi sui residui siano rispettate
5. Non interpretare puntualmente i valori dei coefficienti di regressione

## Analisi delle componenti principali

1. Scegliere in maniera non consona il trattamento dei dati: es., standardizzare o non standardizzare? Trasformare tutte o alcune variabili oppure no?
2. Non adottare o non applicare correttamente un criterio quantitativo (percentuale di varianza spiegata, autovalori maggiori della loro media, etc.) per la scelta del numero di componenti principali, dimensione del sottospazio in cui si proiettano le variabili
3. Non interpretare affatto, o interpretare arbitrariamente le variabili componenti principali e il loro legame con le variabili iniziali

4. Applicare l'analisi delle componenti principali quando una riduzione delle variabili risulta inutile (anzi dannosa, in quanto informazione potenzialmente utile non viene presa in considerazione), ad esempio se  $n$  e  $p$  sono entrambi piccoli, etc.

## Analisi dei cluster

1. Utilizzare variabili altamente correlate
2. Rendere quantitative variabili qualitative in modo arbitrario e non logico (non coerente)
3. Non ridurre le variabili mediante componenti principali al fine di applicare una tecnica gerarchica per la determinazione endogena del numero di cluster, oppure applicare una tecnica non gerarchica con una numerosità dei cluster arbitraria (non suggerita dal problema), senza premettere un'analisi gerarchica sulle componenti principali
4. Non interpretare correttamente i singoli cluster mediante il confronto di sintesi (min, max, media, varianza, deviazione standard, asimmetria, curtosi) e quantili delle distribuzioni delle variabili

iniziali, nella popolazione e nelle sottopopolazioni ottenute da ciascun cluster, oppure interpretarli su sintesi e quantili delle variabili componenti principali in vece che su quelle riferite alle variabili iniziali (confronto su un sottoinsieme dell'informazione disponibile, perdita quindi di informazioni)

5. Non specificare le caratteristiche dell'algoritmo applicato, tipo di legame, di distanza scelta, etc.
6. Applicare un solo metodo (o gerarchico, o non gerarchico), in vece che applicarne uno gerarchico e uno non gerarchico (ricerca incompleta)
7. Non operare un confronto critico fra la suddivisione in cluster ottenuta con il metodo gerarchico e quella ottenuta con il metodo non gerarchico

## Analisi discriminante

1. Non interpretare le (oppure non dare un significato ragionevole alle) funzioni discriminanti, in relazione alla loro correlazione e al loro legame con le variabili iniziali

2. Non scegliere in modo opportuno la dimensione del sottospazio delle funzioni discriminanti

### Analisi delle corrispondenze

1. Non scegliere in modo opportuno la dimensione del sottospazio dei fattori
2. Non interpretare i (oppure non dare un significato ragionevole ai) fattori, in relazione alla loro correlazione e al loro legame con le modalità e variabili iniziali
3. Non interpretare correttamente i grafici, nel caso di proiezioni in sottospazi di dimensione 2 o 3
4. Omettere, o considerare parzialmente i contributi puntuali e globali di modalità e fattori.