# Statistica Inferenziale per Data Scientist

## Progetto Finale

Per questo studio medico si analizzano i dati raccolti da 3 ospedali, riguardanti 2500 neonati.

In particolare si sono rilevate le seguenti variabili:

* età della madre
* numero di gravidanze sostenute
* Madre fumatrice (0=NO, SI=1)
* N° di settimane di gestazione
* peso in grammi del neonato
* Lunghezza in mm del neonato
* Diametro in mm del cranio del neonato
* Tipo di parto: Naturale o Cesareo
* Ospedale: 1, 2, 3
* Sesso del neonato: M o F

Si vuole scoprire se è possibile prevedere il peso del neonato alla nascita date tutte le altre variabili.

In particolare, si vuole studiare una relazione con le variabili della madre, per capire se queste hanno o meno un effetto significativo, come ad esempio l’effetto potenzialmente dannoso del fumo (nascite premature?). Si usano anche lunghezza e diametro del cranio del neonato perché si possono stimare già dalle ecografie, ma in generale potrebbero anche fungere da variabili di controllo).

Puoi seguire i punti che ti scrivo io come traccia e svolgerli uno alla volta, commentando ovviamente i risultati.

1. Importa il dataset “neonati.csv” e controlla che sia stato letto correttamente dal software
2. Descrivi il dataset, la sua composizione, il tipo di variabili e l’obbiettivo dello studio
3. Indaga le variabili effettuando una breve analisi descrittiva, utilizzando indici e strumenti grafici che conosci
4. Saggia l’ipotesi che la media del peso e della lunghezza di questo campione di neonati siano significativamente uguali a quelle della popolazione
5. Per le stesse variabili, o per altre per le quali ha senso farlo, verifica differenze significative tra i due sessi
6. Si vocifera che in alcuni ospedali si facciano più parti cesarei, sai verificare questa ipotesi?

Analisi multidimensionale:

1. Ricordati qual è l’obbiettivo dello studio e indaga le relazioni a due a due, soprattutto con la variabile risposta
2. Crea un modello di regressione lineare multipla con tutte le variabili e commenta i coefficienti e il risultato ottenuto
3. Cerca il modello “migliore”, utilizzando tutti i criteri di selezione che conosci e spiegali.
4. Si potrebbero considerare interazioni o effetti non lineari?
5. Effettua una diagnostica approfondita dei residui del modello e di potenziali valori influenti. Se ne trovi prova a verificare la loro effettiva influenza
6. Quanto ti sembra buono il modello per fare previsioni?
7. Fai la tua migliore previsione per il peso di una neonata, considerato che la madre è alla terza gravidanza e partorirà alla 39esima settimana. Niente misure dall’ecografia.
8. Cerca di creare qualche rappresentazione grafica che aiuti a visualizzare il modello. Se è il caso semplifica quest’ultimo!

Produci un documento di testo o pdf in cui riporti il codice, i grafici, e commenti spiegando ciò che hai fatto.

Consegna anche l’Rscript (il file.R). Puoi consegnare anche solo quest’ultimo, scrivendo i commenti sopra le righe di codice e ovviamente il punto corrispondente.

Altre considerazioni e analisi fuori dagli schemi sono assolutamente ben accette!

Se hai problemi usa il manuale integrato di R utilizzando la funzione “help(nomefunzione)” oppure “?nomefunzione”, per capire come usare tutti i vari argomenti… e non dimenticarti che Google è tuo amico, anzi, è pieno di amici!

Non c’è alcun limite imposto per l’utilizzo di R e la fantasia! Se sei uno smanettone come me, divertiti a cercare pacchetti e funzioni che possano agevolarti il lavoro.

Se non sei un principiante della programmazione, sbizzarrisciti pure!

## Consegna del progetto

Invia R script e il documento a [**assignments@profession.ai**](mailto:assignments@profession.ai)**, la correzione richiede tra i 7 e i 14 giorni lavorativi.**