Data Science Lab in Medicine
Big Data in Health Care

Master Degree in
DATASCIENCE
University of Milano-Bicocca

Prova finale parte prima: R lab

Prof.ssa Maria Grazia Valsecchi
Dott. Davide Paolo Bernasconi
Dott.ssa Giulia Capitoli

BICOCCA BIOINFORMATICS BIOSTATISTICS AND BIOIMAGING CENTER – B4
SCHOOL OF MEDICINE AND SURGERY, UNIVERSITY OF MILANO-BICOCCA

Assegnazione e dei due lab

- **Studenti con matricola che finisce con numero PARI (o ZERO):**
  - **LAB 1**: valutazione della **performance predittiva di un modello** di rischio con covariate al basale e confronto con un **modello «aumentato»** ovvero che include gli stessi predittori del primo con l'aggiunta di un **ulteriore marker.**
    DATASET: valvola aortica

- **Studenti con matricola che finisce con numero DISPARI:**
  - **LAB 2**: costruzione di un modello di rischio con **covariate cliniche** e sviluppo di un modello «aumentato» ovvero che include tra i predittori anche **un set di variabili di espressione genica selezionate tramite regressione penalizzata.**
    DATASET: breast cancer

DATASET: valvola aortica

Studio osservazionale sulla sopravvivenza dei pazienti che hanno subìto l'impianto di valvole cardiache. Variabili incluse nel dataset:

- Paz.id: identificativo del paziente.
- log.lvmi: logaritmo natural dell'indice di massa ventricolare sinistra "Left Ventricular Mass Index" (standardizzato) misurato al basale.
- fuyrs: tempo di follow-up dalla chirurgia (in anni).
- status: indicatore di evento (1 = morto; 0 = perso al follow up).
- sex: genere del paziente (0 = M; 1 = F).
- age: età del paziente (in anni) alla chirurgia.
- con.cabg: presenza concomitante di bypass coronarico (1 = si; 0 = no).
- creat: creatinina serica pre-operatoria (µmol/mL)
- lv: frazione di eiezione ventricolare sinistra pre-operatoria (1 = buona, 2 = moderata, 3 = scarsa).
- sten.reg.mix: emodinamica della valvola aortica (1 = stenosi, 2 = rigurgito, 3 = misto).

DATASET: breast cancer

Studio osservazionale sulla sopravvivenza libera da metastasi in donne affette da cancro al seno con coinvolgimento linfonodale. Variabili incluse nel dataset:

- time: tempo di follow-up libero da metastasi (in mesi).
- event: indicatore di evento (1 = metastasi o morte; 0 = censura).
- Diam: Diametro del tumore (2 livelli).
- N: numero di linfonodi coinvolti (2 livelli).
- ER: status del recettore di estrogeni (2 livelli).
- Grade: grado del tumore (tre livelli ordinati).
- Age: età della paziente alla diagnosi (in anni).
- TSPYL5 … C20orf46: misura di espressione genica di 70 geni potenzialmente prognostici

Analisi richieste per LAB 1

- analisi descrittive di tutte le variabili presenti nel dataset.
- analisi univariate (modello di Cox) dell'associazione di ciascuna variabile indipendente con l'outcome in studio.
- sviluppo del modello predittivo «base» (modello di Cox) con covariate: sex, age, con.cabg, creat, lv, sten.reg.mix. Senza includere log.lvmi.
- valutare forma funzionale delle variabili continue e assunzione «Proportional Hazards»
- sviluppo del modello predittivo «aumentato» (modello di Cox) con covariate: sex, age, con.cabg, creat, lv, sten.reg.mix e includendo anche log.lvmi.
- valutare forma funzionale delle variabili continue e assunzione «Proportional Hazards» (in particolare per log.lvmi).
- valutazione e confronto della performance dei due modelli (calibrazione, discriminazione, Net Benefit) per la predizione del rischio di evento ad un time-point fisso (es. 5 anni).
- predizione del rischio di evento ad un time-point fisso (es. 5 anni) per 3 soggetti «tipo» (scelti casualmente nel dataset o nuovi ipotetici soggetti) basata sul modello base e sul modello aumentato.

Analisi richieste per LAB 2

- analisi descrittive di tutte le variabili presenti nel dataset.
- analisi univariate (modello di Cox) dell'associazione di ciascuna variabile clinica (Diam, N, ER, Grade, Age) con l'outcome in studio.
- sviluppo del modello predittivo «base» (modello di Cox) con covariate cliniche: Diam, N, ER, Grade, Age. Non includere le variabili di espressione genica.
- valutare forma funzionale delle variabili continue e assunzione «Proportional Hazards»
- analisi univariate (modello di Cox) dell'associazione di ciascuna di espressione genica con l'outcome in studio. Selezionare le variabili con associazione significativa ($p<0.05$) dopo aggiustamento con metodo di Benjamini-Hochberg.
- selezione delle variabili di espressione genica associate con l'outcome tramite modello di Cox penalizzato con metodo LASSO. Non includere nel modello le variabili cliniche.
- sviluppo del modello predittivo «aumentato» (modello di Cox) ovvero con variabili cliniche e variabili di espressione genica selezionate al punto precedente
- predizione del rischio di evento ad un time-point fisso (es. 1 anno) per 3 soggetti «tipo» (scelti casualmente nel dataset o nuovi ipotetici soggetti) basata sul modello base e sul modello aumentato

Da consegnare (per entrambi i LAB)

1. Report dei risultati
   - in formato .pdf o .docx o .odt o .html
   - Mostrare l'output dell'analisi (grafici e tabelle)
   - Includere un breve commento discorsivo ad ogni risultato chiarendone l'interpretazione
   - Aggiungere una piccola discussione finale, ovvero riassumere i risultati e menzionare eventuali limiti dei dati e problematiche riscontrate durante analisi

2. Codice R
   - in formato .R (oppure .Rmd)

Date ed esito

- Caricare il progetto sulla piattaforma e-learning (nella apposita sezione) **entro il 30 maggio.**

- La valutazione del progetto peserà per il 60% sul voto finale. Il quiz a risposta multipla peserà quindi per il 40%.

- Per ulteriori informazioni o chiarimenti utilizzare il forum della pagina e-learning del corso o scrivere a davide.bernasconi@unimib.it