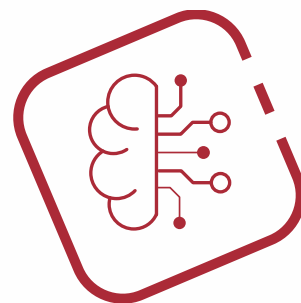


# ESAME

Dott. Francesco Stranieri

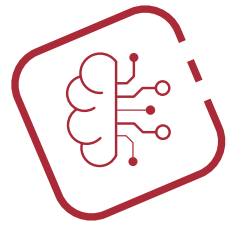


# ESAME

Heart Disease

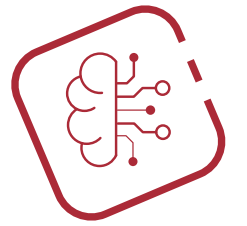
[https://archive.ics.uci.edu/  
ml/datasets/Heart+Disease](https://archive.ics.uci.edu/ml/datasets/Heart+Disease)





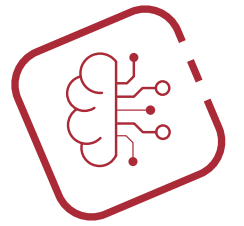
# Consegna

1. Caricare il dataset `heart.csv` e analizzarne dettagliatamente la struttura.
2. Trasformare i dati in modo che siano *tecnicamente corretti*.
3. Rinominare le colonne in maniera appropriata e descrivere il *tipo* di ogni attributo (nominale, ordinale, di intervallo o di rapporto).
4. Rinominare e ordinare i *livelli* dei fattori in maniera appropriata, se necessario.
5. Descrivere brevemente gli attributi.



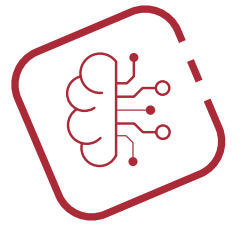
# Consegna

1. Controllare se sono presenti valori NA e, nel caso, rimuoverli.
2. Rimuovere le colonne ritenute non necessarie, se presenti.



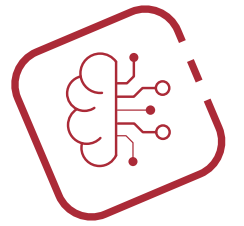
# Consegna

1. Trasformare i dati in modo che siano *consistenti*. Assumere, ad esempio, che la frequenza cardiaca massima non possa essere superiore a 222, sostituendo i valori maggiori di 222 con il valore medio della variabile.
2. Trasformare i dati in modo che siano *consistenti*. Assumere come outlier, ad esempio, i valori relativi alla pressione sanguigna a riposo che non rispettano la *1.5xIQR Rule*. Individuare e rimuovere tali valori.
3. Trasformare i dati in modo che siano *consistenti*. Quali altre trasformazioni sono necessarie?
4. Visualizzare, prima e dopo le trasformazioni, i *grafici* ritenuti più opportuni.



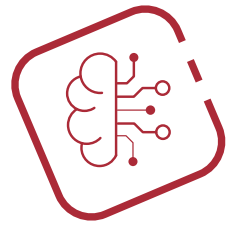
# Consegna

1. Condurre una *analisi descrittiva* approfondita.



# Consegna

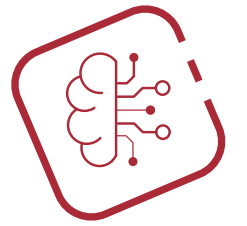
1. Analizzare la relazione tra due variabili (appropriate) del dataset attraverso la *regressione lineare semplice* e determinare:
  - il coefficiente angolare e l'intercetta (anche *interpretabile*) della retta di regressione;
  - il grafico del modello;
  - il tipo di relazione tramite  $r$  e la bontà del modello tramite  $R^2$ ;
  - l'analisi dei residui e la distribuzione in quantili, con i relativi grafici.
2. Creare un data frame contenente 10 osservazioni (non presenti nel dataset) ed effettuare delle *previsioni*.



# Consegna

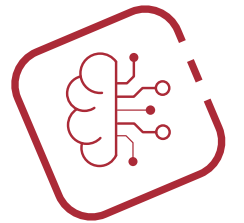
1. Utilizzando le stesse due variabili del dataset scelte per la *regressione lineare semplice*, determinare attraverso l'*algoritmo di forza bruta* (circa 15 mila combinazioni):
  - il coefficiente angolare e l'intercetta della retta di regressione;
  - il grafico del modello.
2. Utilizzando le stesse due variabili del dataset scelte per la *regressione lineare semplice*, determinare attraverso l'*algoritmo del gradiente*:
  - il coefficiente angolare e l'intercetta della retta di regressione;
  - il grafico del modello.
3. Determinare il tempo necessario per calcolare la soluzione (ad esempio, utilizzando la funzione `Sys.time()`) attraverso la *regressione lineare semplice*, l'*algoritmo di forza bruta* e l'*algoritmo del gradiente*, commentando le eventuali differenze.
4. **FACOLTATIVO** Determinare la memoria necessaria per calcolare la soluzione (ad esempio, utilizzando la libreria `profmem`) attraverso la *regressione lineare semplice*, l'*algoritmo di forza bruta* e l'*algoritmo del gradiente*, commentando le eventuali differenze.





# Consegna

1. Applicare un *modello di Machine Learning* a scelta, misurandone l'*accuratezza* sul test set.
2. Descrivere brevemente il funzionamento del modello scelto.



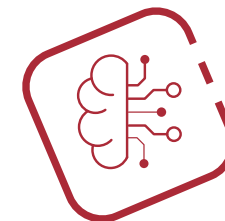
# Valutazione finale

## Progetto (2-3 persone)

- Codice (in R)
  - Il codice sorgente deve essere ben documentato, ben strutturato e seguire le [guide di stile](#)
- Report (5-15 pagine, in PDF)
  - Il testo deve essere strutturato e organizzato in modo chiaro e logico
- Slide (12-15 minuti, in PPT)
  - Le slide devono servire da linee guida per il discorso che si vuole intraprendere

## Valutazione (in trentesimi)

- 60% Codice e Report, 40% Discussione
- 1 punto extra per l'utilizzo di `tydiverse` (`tibble` e `dplyr`)



# Valutazione finale

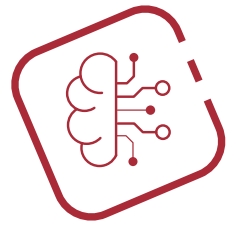
Ogni rappresentante del gruppo deve inviare via mail all'indirizzo [f.stranieri@itsrizzoli.it](mailto:f.stranieri@itsrizzoli.it) un file .zip, denominato «[ML] Cognome1 Cognome2 Cognome3 Gruppo\_{NumeroGruppo} CR», contenente il Codice e il Report.

Termine per la consegna il 27/02/2022 alle 23:59:59.

Ogni rappresentante del gruppo deve inviare via mail all'indirizzo [f.stranieri@itsrizzoli.it](mailto:f.stranieri@itsrizzoli.it) un file .zip, denominato «[ML] Cognome1 Cognome2 Cognome3 Gruppo\_{NumeroGruppo} S», contenente le Slide.

Termine per la consegna il 06/03/2022 alle 23:59:59.

Discussione il 14/03/2022 in presenza all'ITS Rizzoli dalle 08:40 alle 12:40 (alternativamente, il 12/03/2022 su Meet, sempre dalle 08:40 alle 12:40).



# Consigli

- Report (in PDF)
  - Possibili capitoli:
    1. Introduzione e obiettivi
    2. Descrizione del dataset
    3. Analisi dei dati (tecnicamente corretti e consistenti)
    4. Analisi descrittiva
    5. Regressione lineare
    6. Machine Learning
    7. Conclusioni

E' fortemente consigliato l'uso di `git` per la collaborazione e l'uso di `Overleaf` (LaTeX) per la stesura del Report.

**Importantissimo giustificare le scelte prese, spiegando il ragionamento adottato!**

**In caso di copiatura, il progetto verrà valutato come insufficiente.**

In caso di dubbi o problemi potete contattarmi via mail all'indirizzo [f.stranieri@itsrizzoli.it](mailto:f.stranieri@itsrizzoli.it).