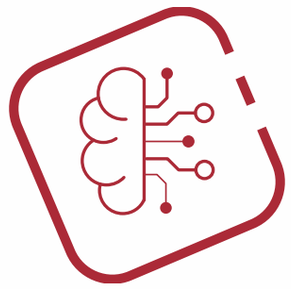
ITSAR ISTITUTO TECNICO SUPERIORE
ANGELO RIZZOLI

ESAME

Francesco Stranieri



ITSAR

ESAME

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 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.

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*. Sono necessarie altre trasformazioni? Se si, quali?
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 del dataset attraverso la *regressione lineare semplice e* determinare:
   - il grafico del modello;
   - il coefficiente angolare e l'intercetta (*interpretabile*) della retta di regressione;
   - 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 (<u>non</u> presenti nel dataset) ed effettuare delle *previsioni*.

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 (1 o 2 persone)

- Codice (in R)
  - Il codice sorgente deve essere ben strutturato, commentato e seguire le guide di stile
- Report (in PDF)
  - Il testo deve essere strutturato e organizzato in modo chiaro e logico

Valutazione (in trentesimi)

- 60% Codice e 40% Report
- 2 punti extra per l'utilizzo di `tydiverse`
  (1 punto per `tibble` + `dplyr` e 1 punto per `ggplot2`)

Valutazione finale

Ogni studente deve caricare sul proprio Drive un file `.zip`, denominato `cognome_nome_ML`, contenente il Codice e il Report.

Termine per la consegna 28/02/2021 alle 23:59:59.

In caso di dubbi o problemi potete contattarmi via mail all'indirizzo f.stranieri@itsrizzoli.it.

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.