Predizione del Rischio Cardiovascolare

Un approccio essenziale con Regressione Logistica

Prof. Fedeli Massimo

Obiettivo del sistema

L'algoritmo realizza un sistema di **Machine Learning supervisionato** per la predizione del rischio cardiovascolare.
Obiettivo principale:

- classificare ogni paziente come **sano** o **a rischio**

Il modello utilizzato è la **regressione logistica**, scelta per la sua semplicità e interpretabilità.

Dataset clinico

Il sistema utilizza un dataset di pazienti reali contenente:

- dati anagrafici (età, sesso)
- parametri clinici (pressione sanguigna, colesterolo)
- risultati di test diagnostici

I dati sono caricati da una sorgente online e organizzati in un **DataFrame**.

Gestione dei valori mancanti

Nel dataset originale sono presenti valori mancanti indicati con un simbolo speciale.
La procedura adottata è:

- conversione dei simboli in valori nulli
- eliminazione delle osservazioni incomplete

Questa scelta, seppur semplice, è adeguata a una versione didattica dell'algoritmo.

Definizione della variabile target

La variabile di interesse rappresenta il grado di malattia cardiaca.
Nel dataset originale:

- il livello di malattia è espresso con più valori interi

L'algoritmo semplifica il problema trasformando il target in:

- 0 = assenza di malattia
- 1 = presenza di malattia

Classificazione binaria

Questa trasformazione consente di:

- ricondurre il problema a una **classificazione binaria**
- rendere il modello più interpretabile
- utilizzare in modo naturale la regressione logistica

Il focus diventa la stima del rischio, non la gravità clinica.

Preparazione dei dati

Il dataset viene suddiviso in:

- variabili di input (feature cliniche)
- variabile di output (stato di salute)

Successivamente i dati sono separati in:

- training set
- test set

Suddivisione stratificata

La suddivisione del dataset è **stratificata**.
Questo significa che:

- la proporzione tra pazienti sani e malati è mantenuta
- si evitano distorsioni nella valutazione del modello

La valutazione risulta così più affidabile.

Standardizzazione delle variabili

Le variabili cliniche hanno scale molto diverse:

- età in anni
- valori di laboratorio
- indicatori discreti

Prima dell'addestramento viene applicata la **standardizzazione**.

Perché standardizzare

La standardizzazione:

- centra le variabili sulla media
- imposta deviazione standard unitaria
- rende le feature confrontabili tra loro

Questo passaggio è particolarmente importante per la regressione logistica, sensibile alle differenze di scala.

Regressione logistica

La regressione logistica è un algoritmo di classificazione che:

- stima la probabilità di appartenenza alla classe "a rischio"
- utilizza una funzione sigmoide

Il risultato è una probabilità compresa tra 0 e 1.

Addestramento del modello

Durante l'addestramento:

- il modello apprende dai dati di training
- stima un peso per ciascuna variabile clinica

I pesi rappresentano l'influenza di ogni fattore sul rischio cardiovascolare.

Valutazione del modello

Il modello viene testato su pazienti mai visti in precedenza.
Le previsioni sono confrontate con i valori reali per calcolare:

- **accuratezza**

L'accuratezza indica la percentuale di classificazioni corrette.

Considerazioni finali

L'accuratezza fornisce una prima valutazione dell'efficacia del sistema.
Tuttavia:

- non esaurisce tutte le metriche clinicamente rilevanti
- rappresenta un buon punto di partenza didattico

Il modello è semplice, interpretabile e adatto a scopi formativi.