

Sistema di Predizione del Rischio Cardiovascolare

Algoritmo di Regressione Logistica

Prof. Fedeli Massimo

IIS Fermi Sacconi Cria di Ascoli Piceno

Tutti i diritti riservati

Obiettivo del sistema

Il sistema realizza una predizione del rischio cardiovascolare basata su tecniche di Machine Learning supervisionato.

L'obiettivo principale è **classificare** ciascun paziente in una delle due categorie:

- paziente sano
- paziente a rischio di malattia cardiaca

La classificazione viene effettuata tramite un modello di regressione logistica addestrato su dati clinici reali.

Dataset e caratteristiche

Il punto di partenza è il caricamento di un **dataset** contenente informazioni cliniche e anagrafiche dei pazienti.

Tra le principali variabili considerate:

- età
- sesso
- pressione sanguigna
- colesterolo
- risultati di test diagnostici

I dati vengono letti direttamente da una sorgente online e organizzati in un DataFrame.

Pulizia e trasformazione dei dati

Nel dataset originale sono presenti valori mancanti, indicati tramite un simbolo speciale.

La fase di **preprocessing** prevede:

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

Questa scelta, pur semplice, risulta adeguata per una versione didattica dell'algoritmo.

Definizione del target

Nel dataset originale il grado di malattia è espresso mediante più valori interi (0,1,2,3,4).

Il problema viene semplificato trasformando il target in forma binaria:

- assenza di malattia (0)
- presenza di malattia (1)

Questa trasformazione rende il problema più adatto alla regressione logistica e più semplice da interpretare.

Suddivisione del dataset

La preparazione dei dati prevede la separazione tra:

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

Il dataset viene suddiviso in:

- insieme di addestramento
- insieme di test

La suddivisione è stratificata, per mantenere proporzioni simili di pazienti sani e malati.

Standardizzazione delle feature

Le variabili cliniche presentano scale molto diverse tra loro.

Prima dell'addestramento viene applicata la standardizzazione:

- media nulla
- deviazione standard unitaria

Questo passaggio è fondamentale per la **regressione logistica**, che è sensibile alle differenze di scala tra le variabili.

Addestramento del modello

Il modello utilizzato è una regressione logistica per problemi di classificazione binaria.

Durante l'addestramento il modello:

- apprende un insieme di pesi
- quantifica l'influenza di ciascuna variabile clinica

L'output del modello è la probabilità che un paziente appartenga alla classe "a rischio".

Valutazione delle prestazioni

Il modello viene valutato sui **dati di test**, non utilizzati durante l'addestramento.

Le previsioni vengono confrontate con i valori reali per calcolare l'accuratezza.

L'accuratezza rappresenta la percentuale di classificazioni corrette e fornisce una prima indicazione dell'efficacia del sistema, pur non esaurendo tutte le possibili metriche clinicamente rilevanti.