Regressione Lineare

Come suggerisce il nome, la regressione lineare è un modello che risolve un problema di regressione, ovvero dato un vettore $\boldsymbol{x} \in \mathbb{R}^n$ in input, restituisce un valore $y \in \mathbb{R}$ in output. L'output della regressione lineare è una funzione lineare dell'input.

Definiamo $\hat{y}$ come il valore che il nostro modello predice, definiamo dunque l'output come:

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}$$

Dove: $\boldsymbol{w}$ è un vettore di parametri.

Questi parametri, anche chiamati pesi, determinano il comportamento del sistema; in questo specifico caso si tratta del coefficiente per cui moltiplichiamo il vettore di input $\boldsymbol{x}$.

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x} + b$$

Questa è una *affine function*, ovvero una funzione lineare con una traslazione (b è noto come *intercept term* o *bias*). Come si può notare, inoltre, l'equazione assomiglia molto a quella di una retta in due dimensioni: $y = mx + q$. Infatti per un grado $n = 1$ la regressione lineare è proprio una retta.

Facciamo un breve esempio pratico: supponiamo di avere un dataset con GDP per capita e un valore di soddisfazione della vita per ogni paese del mondo e volessimo costruire un modello che preveda quest'ultimo valore[1].

Prima di tutto plottiamo i dati:

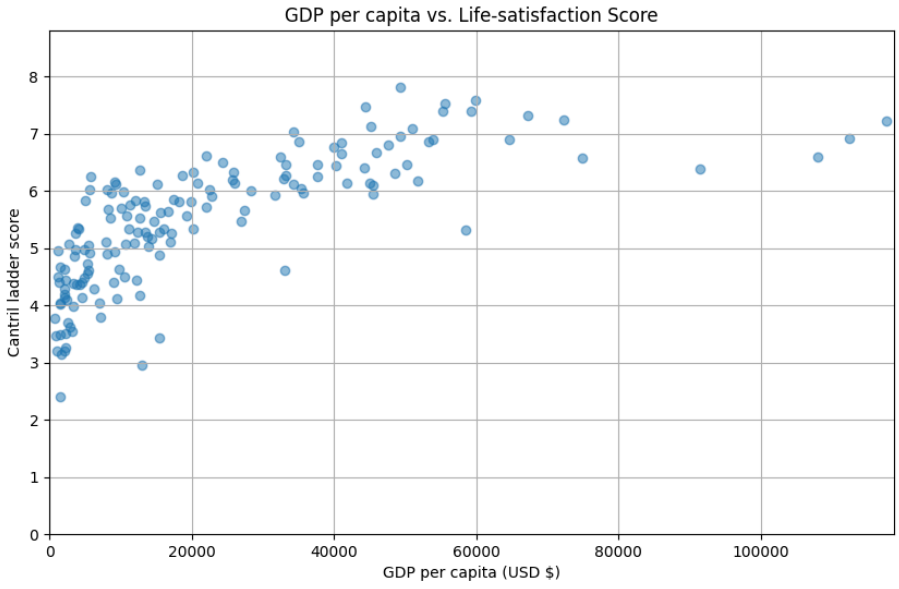

Figure 1: Plot dei dati GDP vs Life-satisfaction degli ultimi dati disponibili per ogni paese. (ex Austria)

Ora proviamo ad utilizzare la regressione lineare per prevedere il livello di soddisfazione della vita in Austria, che abbiamo escluso dal training set, dato il suo GDP per capita:

[1] Questo valore viene misurato con la scala di Cantril.

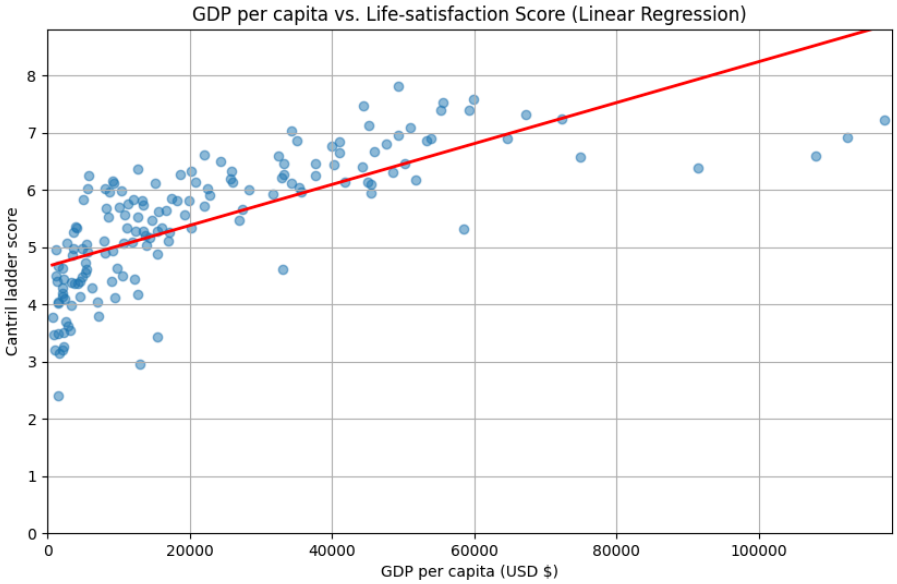

Figure 2: Plot dei dati GDP vs Life-satisfaction con la regressione lineare e grado 1

L'austria nel 2022 aveva un GDP per capita di $55,867 e un livello di felicità di 7,09. Il modello di regressione lineare ci dice che il livello di felicità previsto è di 6,66. Forse possiamo fare di meglio.

Torniamo sulla formula della regressione lineare, possiamo generalizzarla come:

$$\hat{y} = b + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

Dalla formula generalizzata capiamo che la regressione lineare può funzionare anche in più dimensioni, non solo con una variabile indipendente; ed in questo caso si dice "multivariata". Per esempio con 2 variabili indipendenti avremo un piano. Dunque se aggiungessimo lo Human Freedom Index come feature, avremmo un modello tridimensionale:

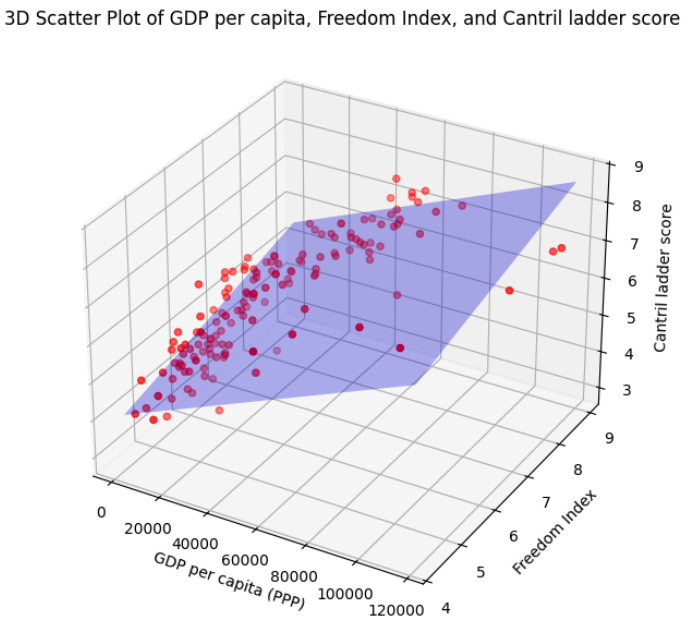

Figure 3: Plot dei dati GDP e Freedom vs Life satisfaction in 3D

In questo caso la predizione del modello per l'Austria è di 6,80, più vicina al valore reale.