REGRESSIONE MULTIPLA E REGRESSIONE LINEARE

A differenza della regressione lineare semplice, la regressione lineare multipla tiene conto dell'effetto combinato di più variabili in ingresso, questo ci consente di verificare la relazione di più valori su una variabile dipendente. La retta di regressione multipla richiede però una più attenta gestione delle variabili indipendenti come vedremo successivamente.

Per svolgere l'Excel è stato preso in considerazione un dataset trovato su internet che tiene conto delle performance degli studenti sotto determinati variabili quali:

1. Hours Studied
2. Previous Scores
3. Sleep Hours
4. Sample Question Papers Practiced
5. Performance Index (Valore target)

L'equazione per la retta di regressione multipla è la seguente:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

Con:

$$\beta_0 = \frac{\sum y \sum x_1^2 x_2^2 x_3^2 x_4^2 - \sum x_1 \sum x_2 \sum x_3 \sum x_4 \sum x_1 x_2 x_3 x_4 y}{n \sum x_1^2 x_2^2 x_3^2 x_4^2 - (\sum x_1)^2 (\sum x_2)^2 (\sum x_3)^2 (\sum x_4)^2}$$

$$\beta_j = \frac{n \sum x_j y - \sum x_j \sum y}{n \sum x_j^2 - (\sum x_j)^2}$$

Calcolati i coefficienti e sostituiti all'equazione, sono stati calcolati i valori di output generate dalla regressione multipla:

Performance M.R
6,16275E+13
4,26502E+13
5,11722E+13
3,83232E+13
5,52241E+13
4,01662E+13
5,38045E+13
4,93417E+13
4,67671E+13
4,41993E+13
6,26237E+13
5,90502E+13
3,00212E+13
4,06292E+13

I valori predetti si discostano molto dai valori reali in quanto, calcolando l'MSE risulta che MSE = 2,2961E+25, cioè l'errore calcolato tra i valori di output previsti dalla retta di regressione e quelli reali, è significativamente grande. Il problema potrebbe essere la scarsa correlazione tra le variabili. Si calcola il coefficiente di Pearson per ciascuna coppia di variabili:

Correlazione tra ore studiate e Perfomance index:	0,37373
Correlazione tra Previous perfomance e Perfomance index:	0,915189
Correlazione tra ore di sonno e Perfomance index:	0,048106
Correlazione tra num. di prove e Perfomance index:	0,043268

Notiamo che c'è una correlazione significativa tra la variabile **Previous perfomance** e la variabile target **Performace_index**, mentre il coefficiente di Pearson calcolato per le altre coppie di features sono basse, indicando una scarsa correlazione**.** Per fare una buona analisi dei dati, si è deciso di utilizzare una regressione lineare tra le due variabili con correlazione più elevata (variabile *previous performace* e variabile target *performace index*), quindi riducendo l'analisi da una regressione lineare multipla ad una regressione lineare semplice.

REGRESSIONE LINEARE

L'equazione della retta di regressione e i relativi parametri sono i seguenti:

$$\hat{y} = ax + b_0$$

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Una volta trovati i coefficienti a e b, sono stati calcolati i valori di output della retta di regressione lineare, sostituendo all'interno dell'equazione al posto x, il vettore corrispondente alla variabile *previous performance*.

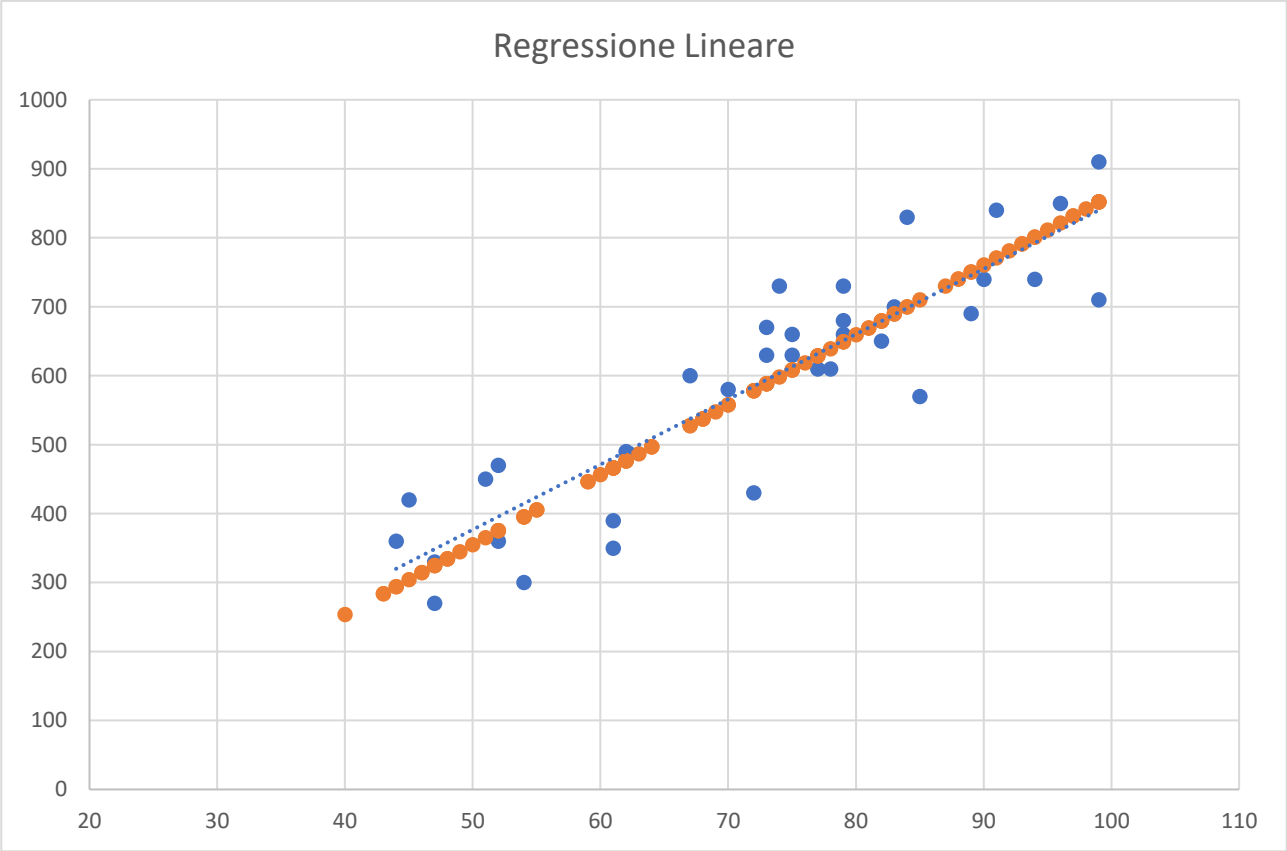
Performance L.R
851,9689134
679,5569781
365,1587432
375,3006217
608,5638283
638,9894639
588,2800712

304,3074719
628,8475854
750,550128
770,833885
649,1313425
324,591229
324,591229

Come previsto i valori predetti dalla retta di regressione, si discostano poco dai valori target. L'MSE misurato su un campione di 100 dati risulta: 61,6088

Che espresso in percentuale rappresenta un errore del 0,61%, decisamente inferiore dell'errore calcolato per la regressione multipla, indicando che il modello di regressione lineare si adatta meglio all'andamento dei dati.

Volendolo rappresentare su un grafico:



(Grafico generato su 100 campioni)

Dove in arancione sono rappresentati i valori generati dal modello di regressione lineare e in blu i valori target.