FONDAMENTI DI SCIENZA DEI DATI: PREDIZIONE DEL PREZZO DI VENDITA TRAMITE REGRESSIONE NON LINEARE

Romano Giovanni, Filippone Emilia, Mazzucco Alex, Cavazza Fabio

Sommario

1.Introduzione

L'obiettivo di questo progetto è analizzare i dati relativi a un Dataset di **Case Ubicate** nel territorio del Boston e prevedere i prezzi tramite un modello di **regressione non lineare**.

Il dataset contiene informazioni di 506 aree urbane, incluse variabili socio-economiche e il valore mediano degli immobili. Per questa analisi, selezioneremo le relazioni di dipendenza tra le variabili descrittive e il prezzo mediano delle abitazioni (**MEDV**).

Il Dataset sarà suddiviso in **due parti**, la prima metà dei dati sarà utilizzata come *training set* per la stima del modello, invece la seconda metà verrà impiegata come *test set* per la valutazione delle prestazioni.

La **performance** del modello sarà valutata tramite **RMSE** (*Root Mean Squared Error*), **MSE** (*Mean Square Error*) e **MAE** (*Mean Absolute Error*) sulla divisione del dataset *testing set* (50%-50%).

2.Descrizione del Dataset

Ogni Entry del Dataset possiede le seguenti features:

Tabella 1 – Descrizione delle features

Sigla	Descrizione
CRIM	Tasso di criminalità pro capite per area
ZN	% di lotti residenziali oltre 25.000 ft²
INDUS	% di superficie non residenziale destinata ad attività industriali
CHAS	Variabile binaria fittizia (=1 se l'area confina con il fiume Charles, 0 altrimenti)
NOX	Concentrazione di ossidi di azoto (parti per dieci milioni)
RM	Numero medio di stanze per abitazione
AGE	% di unità abitative costruite prima del 1940
DIS	Distanza media ponderata ai centri d'impiego
RAD	Indice di accessibilità alle autostrade radiali
TAX	Tasso di imposta sulla proprietà per $10.000
PTRATIO	Rapporto studenti/insegnanti nelle scuole della città
B	Calcolo: $1000(Bk-0.63)^2$, dove Bk è la percentuale di persone afroamericane in città
LSTAT	% di popolazione a basso reddito
MEDV	Valore mediano delle abitazioni occupate dai proprietari (in migliaia di dollari)

Figura 1 – Primi 5 valori del Dataset

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2

Figura 2 - Ultimi 5 valori del Dataset

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
501	0.06263	0.0	11.93	0	0.573	6.593	69.1	2.4786	1	273.0	21.0	391.99	9.67	22.4
502	0.04527	0.0	11.93	0	0.573	6.120	76.7	2.2875	1	273.0	21.0	396.90	9.08	20.6
503	0.06076	0.0	11.93	0	0.573	6.976	91.0	2.1675	1	273.0	21.0	396.90	5.64	23.9
504	0.10959	0.0	11.93	0	0.573	6.794	89.3	2.3889	1	273.0	21.0	393.45	6.48	22.0
505	0.04741	0.0	11.93	0	0.573	6.030	80.8	2.5050	1	273.0	21.0	396.90	7.88	11.9

Figura 2 - Numero di valori mancanti o nulli per ogni feature

```
Shape: (506, 14)
Missing values per column:
 CRIM       0
ZN          0
INDUS       0
CHAS        0
NOX         0
RM          0
AGE         0
DIS         0
RAD         0
TAX         0
PTRATIO     0
B           0
LSTAT       0
MEDV        0
dtype: int64
```

Figura 3 - Describe sul dataset

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	12.653063	22.532806
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	7.141062	9.197104
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000	5.000000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	6.950000	17.025000
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	11.360000	21.200000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	16.955000	25.000000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.970000	50.000000

3.Data Pre-Processing

È essenziale per garantire l'accuratezza del modello di predizione.

Per quel che riguarda la pulizia dei dati, non abbiamo riscontrato la presenza di valori mancanti (come in figura 3), perciò ci siamo concentrati nella ricerca di eventuali informazioni ridondanti riscontrando nessun risultato.

3.1 Studio della correlazione tra le colonne

Per identificare quali feature fossero forti predittori di **MEDV**, abbiamo costruito la ***matrice di correlazione*** di Pearson fra tutte le variabili.

Figura 4 - Tabella delle correlazioni

I risultati più rilevanti sono:

- La correlazione tra **LSTAT** e **MEDV** risulta fortemente *negativa*, con $\rho \approx -0.74$, a indicare che un incremento della percentuale di popolazione a basso reddito è associato a una *diminuzione* del prezzo mediano delle abitazioni.
- La correlazione tra **RM** e **MEDV** è invece positiva con $\rho \approx +0.70$, suggerendo che un maggior numero medio di stanze per abitazione si accompagna a un valore più *elevato* del prezzo mediano delle abitazioni.
- Altre variabili mostrano correlazioni moderate.

Dato che l'obiettivo era costruire un modello a singola variabile, in grado di catturare la relazione più robusta, abbiamo deciso di concentrare la modellazione su **LSTAT** e **RM**.

3.2 Scalatura dei dati

I risultati sono migliori senza scalatura, in questo caso, per tre motivi:

- Dataset specifico - I dati (RM e LSTAT) hanno range non eccessivamente diversi e non contengono valori estremi.
- **Stabilità numerica** - Il dataset di Boston Housing è relativamente piccolo e ben comportato.
- Caratteristiche dei dati – Le relazioni tra le variabili potrebbero essere meglio rappresentate nei loro valori originali.

Abbiamo mantenuto il modello senza scalatura, dato che i risultati sono migliori e le predizioni erano più precise; per gradi polinomiali minori non si notavano miglioramenti, mentre per i gradi più alti si notavano grosse ripercussioni negative. Alla luce di ciò abbiamo preferito **abbandonare** le tecniche di pre-processing.

3.3 Aggregazione e data reduction

Nel dataset Boston Housing ogni osservazione corrisponde a un'unità territoriale distinta ed ogni singola riga contiene l'insieme completo delle informazioni necessarie. Di conseguenza non è stata prevista alcuna operazione di raggruppamento o sintesi aggregata dei dati.

Inoltre, non è stata necessaria alcuna tecnica di *data reduction*, in quanto le dimensioni ridotte del dataset e l'assenza di ridondanze hanno permesso di usare il dataset così com'è.

3.4 Relazioni non lineari delle features

Al fine di investigare la natura delle relazioni tra la variabile dipendente MEDV e le variabili indipendenti RM (numero medio di stanze per abitazione) e LSTAT (percentuale di popolazione a basso reddito), sono stati costruiti e confrontati modelli di regressione lineare e polinomiale di secondo grado, rappresentati graficamente sia in due che in tre dimensioni.

3.4.1 Confronto tra Modello Lineare e Polinomiale (2D)

Nel primo set di grafici (Figura 6), viene evidenziata la relazione tra:

- RM e MEDV: si osserva una *relazione positiva*, **ma non perfettamente lineare**, soprattutto alle estremità del dominio.
- LSTAT e MEDV: la relazione è *marcatamente decrescente e curva*, suggerendo una **chiara non linearità**. Il modello polinomiale (curva verde) si adatta molto meglio rispetto al modello lineare (linea rossa), soprattutto per valori estremi di LSTAT.

3.4.2 Confronto tra Modello Lineare e Polinomiale (3D)

I grafici tridimensionali (Figura 7) rappresentano le superfici di regressione:

- **Il modello lineare** (Grado 1) mostra un piano inclinato che tenta di approssimare la relazione tra LSTAT, RM e MEDV, ma non riesce a catturare la curvatura presente nei dati.
- **Il modello polinomiale** (Grado 2) mostra una superficie più articolata che segue meglio la distribuzione reale dei dati, adattandosi alle variazioni più complesse.

Figura 5 – Relazione lineare tra RM, LSTAT e MEDV

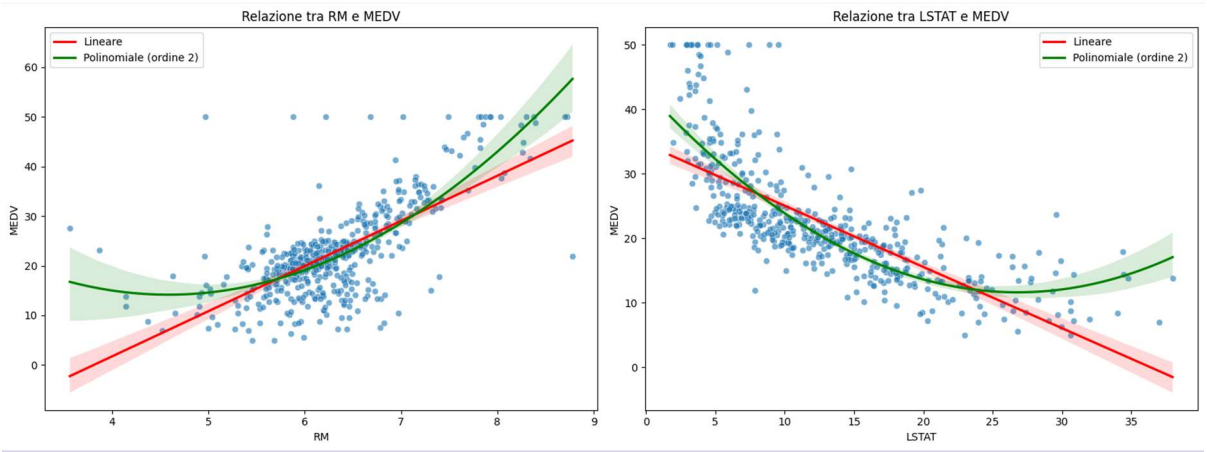
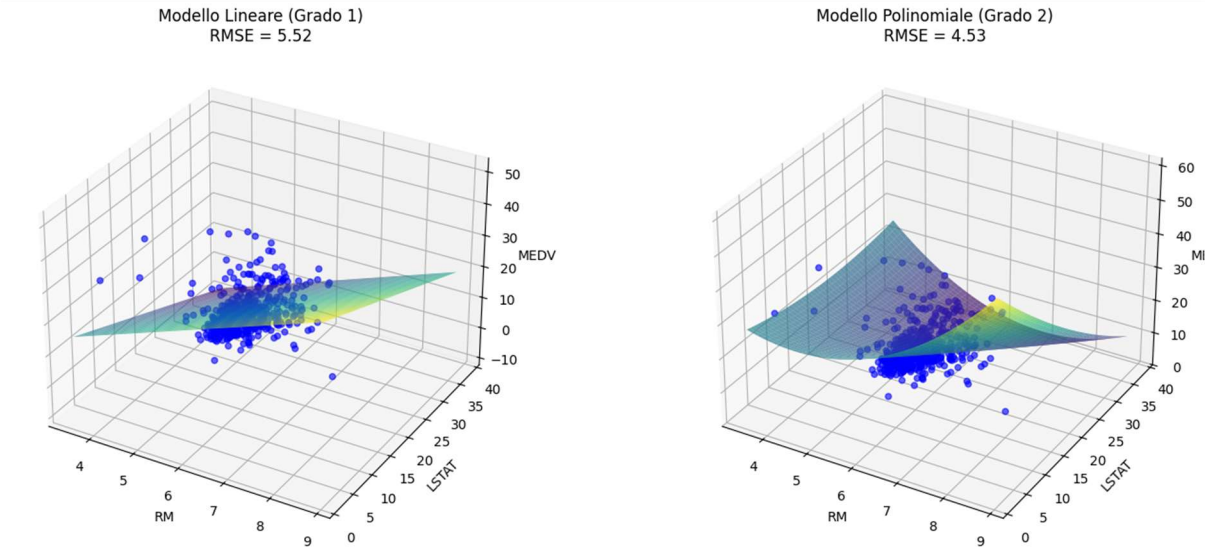


Figura 6 – Grafici rappresentanti l'adattabilità della regressione non lineare rispetto alla lineare



Questi risultati dimostrano che la relazione tra MEDV con RM e LSTAT, non può essere pienamente rappresentata da un modello lineare. L'uso di una regressione polinomiale migliora la qualità dell'adattamento, specialmente nei valori estremi, suggerendo che modelli non-lineari **siano più appropriati** per catturare la vera natura della relazione.

4 Regressione non Lineare

4.1 Studio dei modelli polinomiali

Abbiamo studiato per diversi gradi la funzione di regressione polinomiale, al fine di avere un quadro completo rispetto alle diverse caratteristiche di ogni grado. Infatti, come si vuol notare dalla Tabella 2 e dai grafici in Figura 8, più il grado polinomiale cresce, maggiore sono gli errori calcolati (RMSE, MAE, MSE).

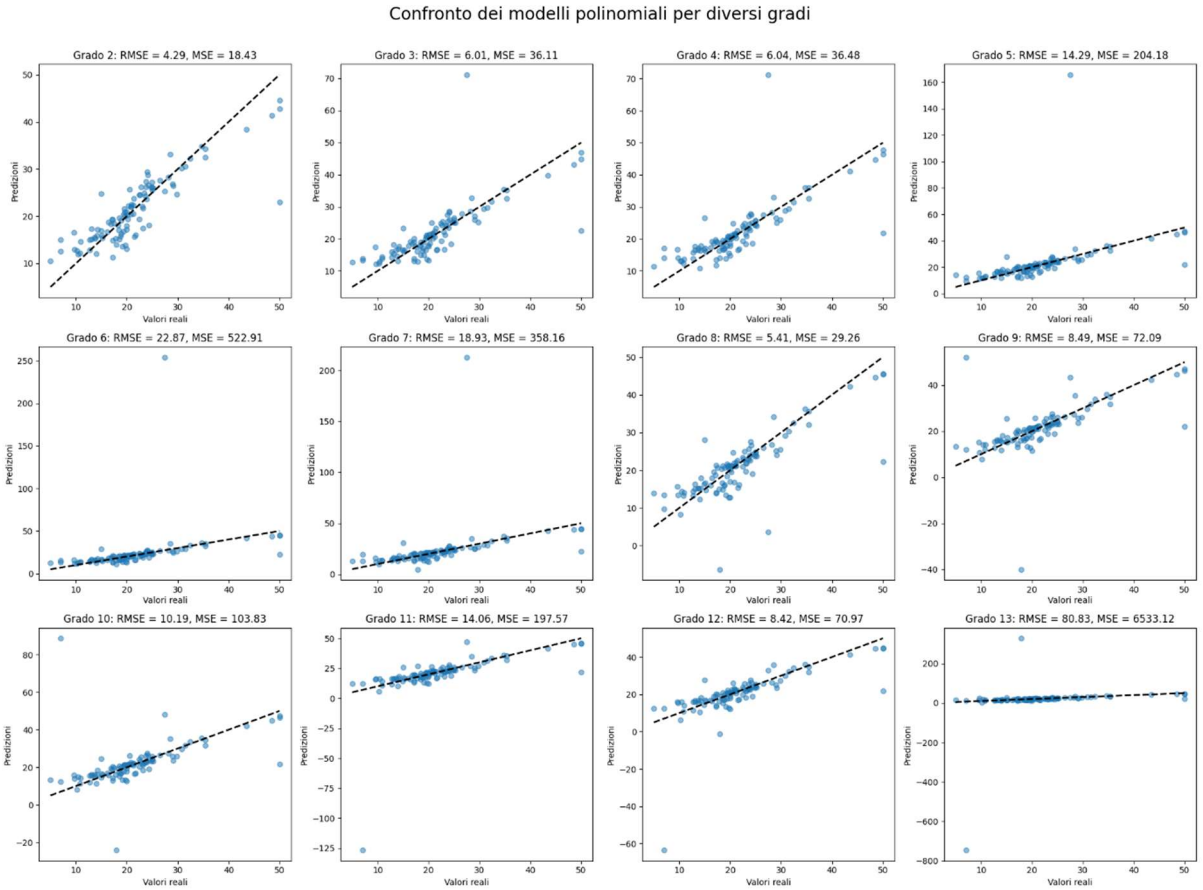
Tabella 2 – Elenco degli errori di ogni grado studiato

	Grado	mse	rmse	mae
0	2	18.433778	4.293458	2.883347
1	3	36.114983	6.009574	3.175291
2	4	36.483673	6.040172	3.092192
3	5	204.180231	14.289165	3.893161
4	6	522.912625	22.867283	4.961402
5	7	358.163113	18.925198	4.685085
6	8	29.260338	5.409283	3.143794
7	9	72.090841	8.490633	3.683711
8	10	103.825327	10.189471	3.988781
9	11	197.574350	14.056114	4.307661
10	12	70.971402	8.424453	3.659355
11	13	6533.118366	80.827708	13.260899

Come conseguenza dell'aumento dei valori degli indici di errore, si nota dalla Figura 8 che i valori predetti si allontanano sempre di più da quelli reali. Per poter leggere al meglio questi grafici, bisogna ricordare che:

- **RMSE (Root Mean Squared Error):** Radice quadrata dell'errore quadratico medio, misura, in media, la differenza tra le predizioni e i valori veri.
- **MSE (Mean Squared Error):** Media dei quadrati degli errori, misura, in media, la distanza tra le predizioni e i valori veri (penalizza maggiormente gli errori più grandi).
- **Asse X:** Indica i valori reali.
- **Asse Y:** Indica i valori predetti dal modello.
- **Punti blu:** Sono le coppie (valore reale e predetto) per ogni osservazione;
- **Linea tratteggiata:** Rappresenta la retta ideale (y = x) dove il modello predice più o meno il valore vero.

Figura 7 – Grafici dei modelli polinomiali dei diversi gradi



Modelli con grado basso (2 - 4)

- Gli errori sono relativamente contenuti.
- Le predizioni seguono abbastanza bene la linea ideale.
- Modello semplice, ma buono.

Modelli con grado medio (5 – 9)

- Gli errori aumentano specialmente ai gradi 5, 6, 7.
- Alcuni punti sono fuori scala rispetto alla linea ideale.
- Inizia pure a comparire l'**overfitting:** il modello si adatta troppo ai dati di addestramento perdendo la capacità di generalizzazione.

Modelli con grado elevato (10 – 13)

- Gli errori diventano molto grandi
- Le predizioni sono distorte, molte completamente fuori scala
- L'overfitting diventa estremo.

I modelli del Grado 2 e 8 offrono i migliori risultati senza troppa complessità

Figura 8 – Rappresentazione degli errori di ogni grado

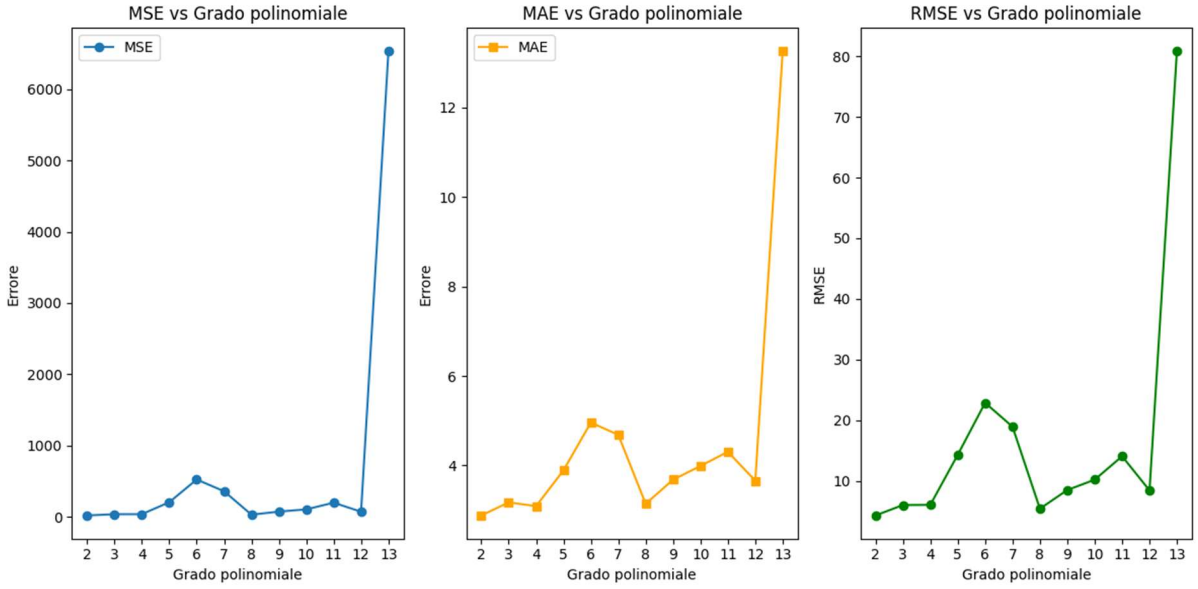
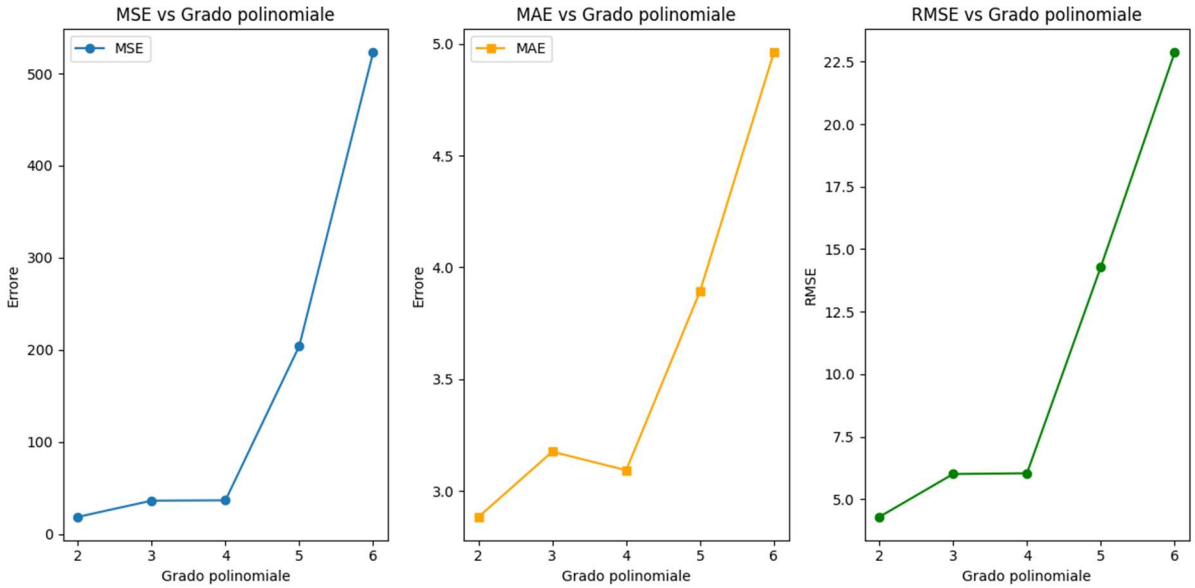


Figura 9 – Zoom-in sui primi 5 gradi



4.2 Predizioni del polinomio migliore

Alla luce di ciò abbiamo scelto il polinomio di grado 2, cioè quello con gli indici di errore più piccoli dei casi studiati. Ciò ha portato a predizioni soddisfacenti.

Figura 10 – 3D scatter plot del modello con il grado di errore minore

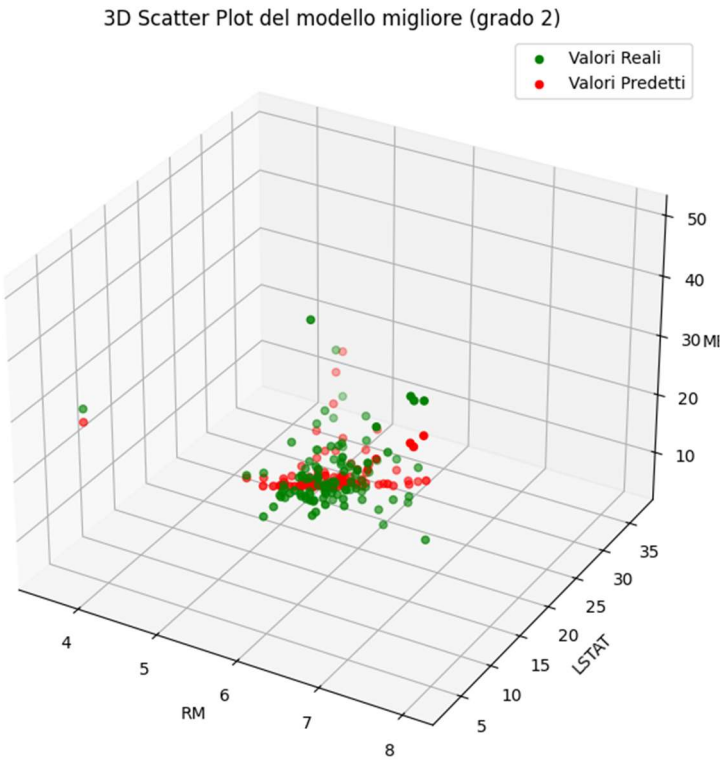


Figura 11 – Grafico 2D del confronto dei valori reali con quelli predetti dal punto di vista di LSTAT

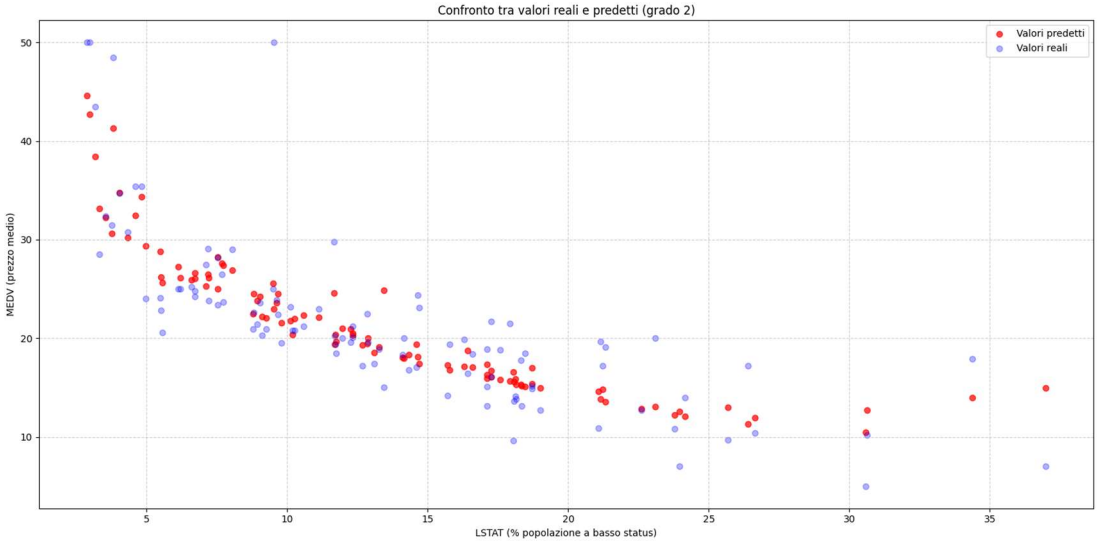
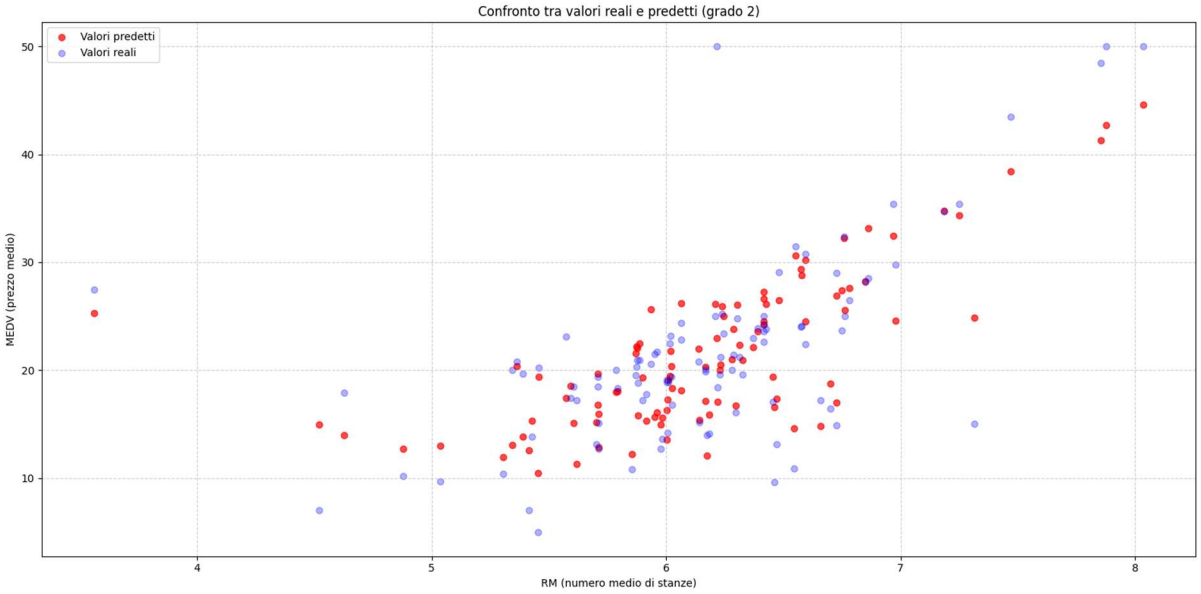


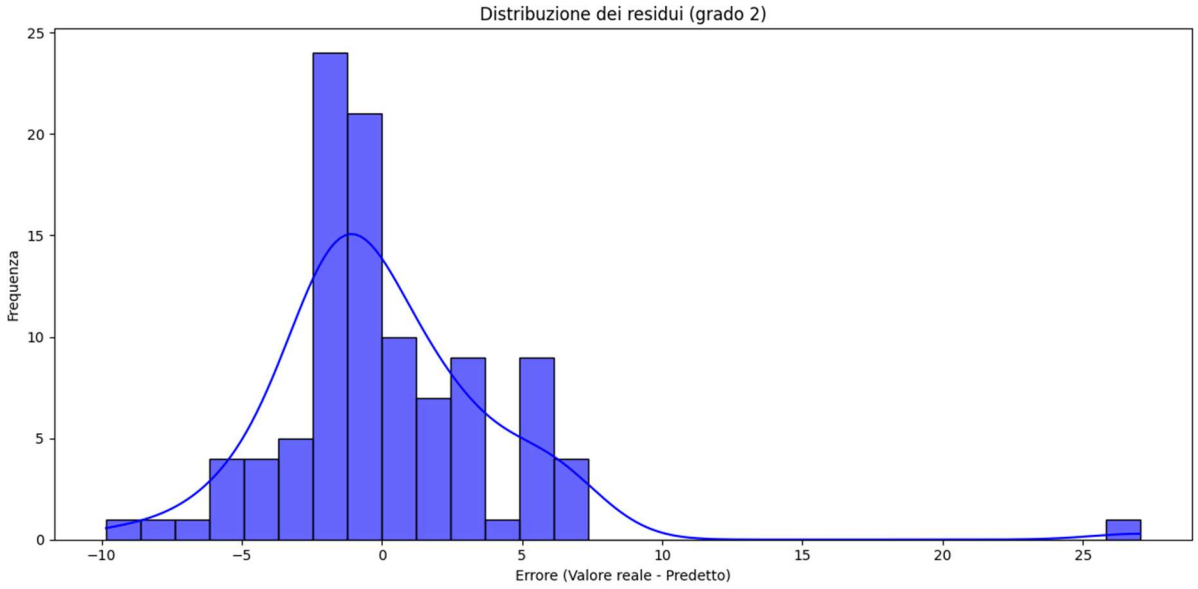
Figura 12- Grafico 2D del confronto dei valori reali con quelli predetti dal punto di vista di RM



4.3 Calcolo dei Residui

I residui, definiti come la differenza tra i valori osservati e quelli stimati dal modello, sono stati analizzati per valutare l'accuratezza delle predizioni. Dall'analisi risulta che il modello tende generalmente a sottostimare il valore reale delle abitazioni, con una differenza media compresa tra i mille e i duemila dollari, come evidenziato nel grafico riportato in Figura 14.

Figura 13 – Studio dei residui



4.4 Predizioni prova

Abbiamo voluto testare le predizioni con dei dati di prova e abbiamo avuto i seguenti risultati (Tabella 3).

Tabella 3 – Tabella riassuntiva dei valori studiati

Risultato delle predizioni (prezzi medi delle case):

	RM	LSTAT	MEDV_predetto
0	5.0	5.0	23.715879
1	6.0	10.0	21.814410
2	7.0	15.0	21.514182
3	8.0	20.0	22.815196
4	5.5	30.0	10.009768
5	6.5	3.0	30.871090

Figura 14 – Grafici 2D dei valori predetti

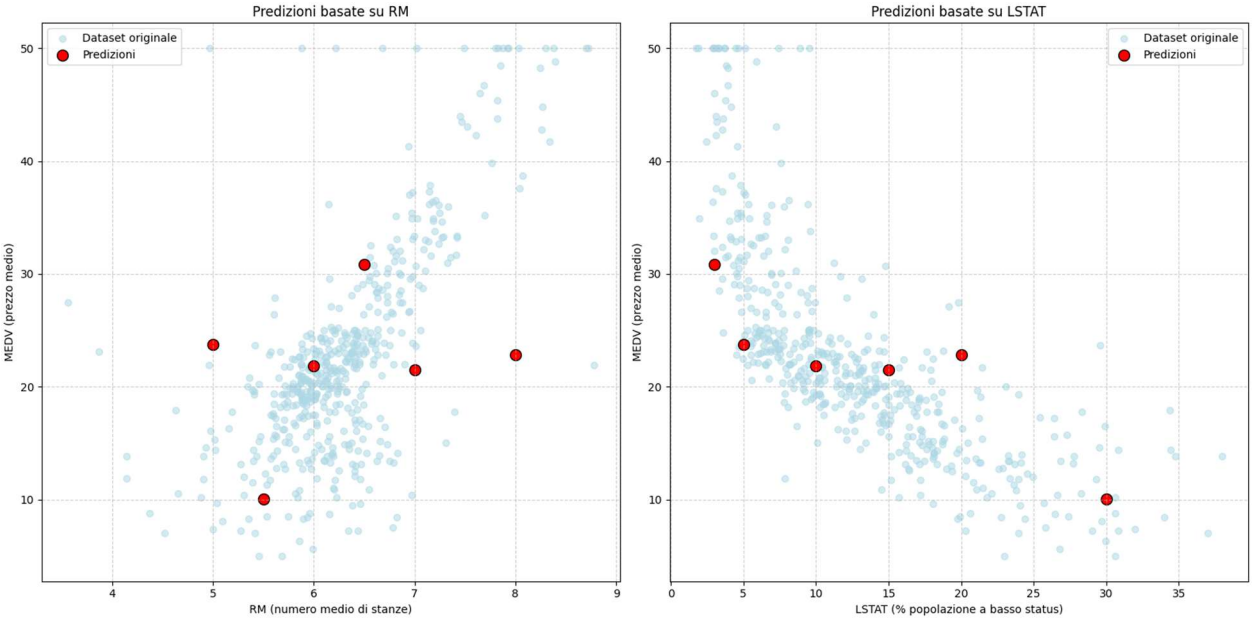
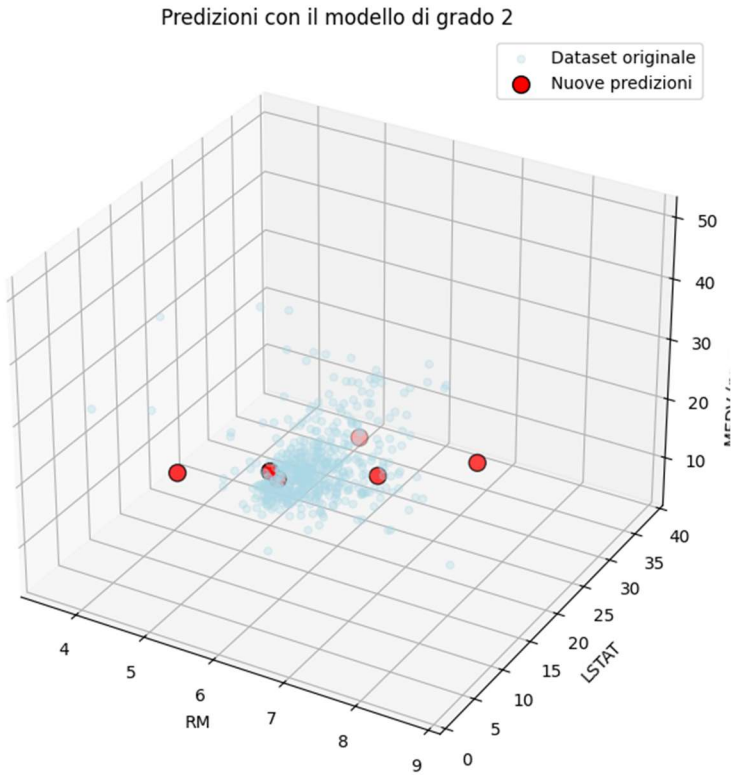


Figura 15 – Grafico 3D dei valori predetti



5 Conclusioni

In conclusione, l'analisi condotta ha messo in evidenza l'influenza significativa di alcune variabili, in particolare il numero medio di stanze per abitazione (RM) e la percentuale di popolazione a basso reddito (LSTAT) e il valore mediano delle case (MEDV). È emerso chiaramente che un maggior numero di stanze è generalmente associato a un aumento del valore degli immobili, mentre un'elevata presenza di popolazione a basso reddito tende a incidere negativamente sul prezzo. Attraverso la matrice di correlazione e il supporto dei grafici, è stato possibile osservare e interpretare con maggiore chiarezza le relazioni tra le variabili. È interessante notare che non sempre una correlazione alta con il target significa che quella variabile migliorerà per forza il modello. Questo può dipendere dal fatto che alcune variabili si somigliano troppo tra loro o che i dati non sono distribuiti in modo perfetto.