CORRIERE DELLA SERA

UNIVERSITÀ DI SIENA 1240

DIISM

CINECA

DISCLAIMER

ULTIME AVVERTENZE PRIMA DELLA RIVOLUZIONE

BIOMIND

INTELLIGENZA ARTIFICIALE PER L'INNOVAZIONE IN FARMACIA E BIOINFORMATICA

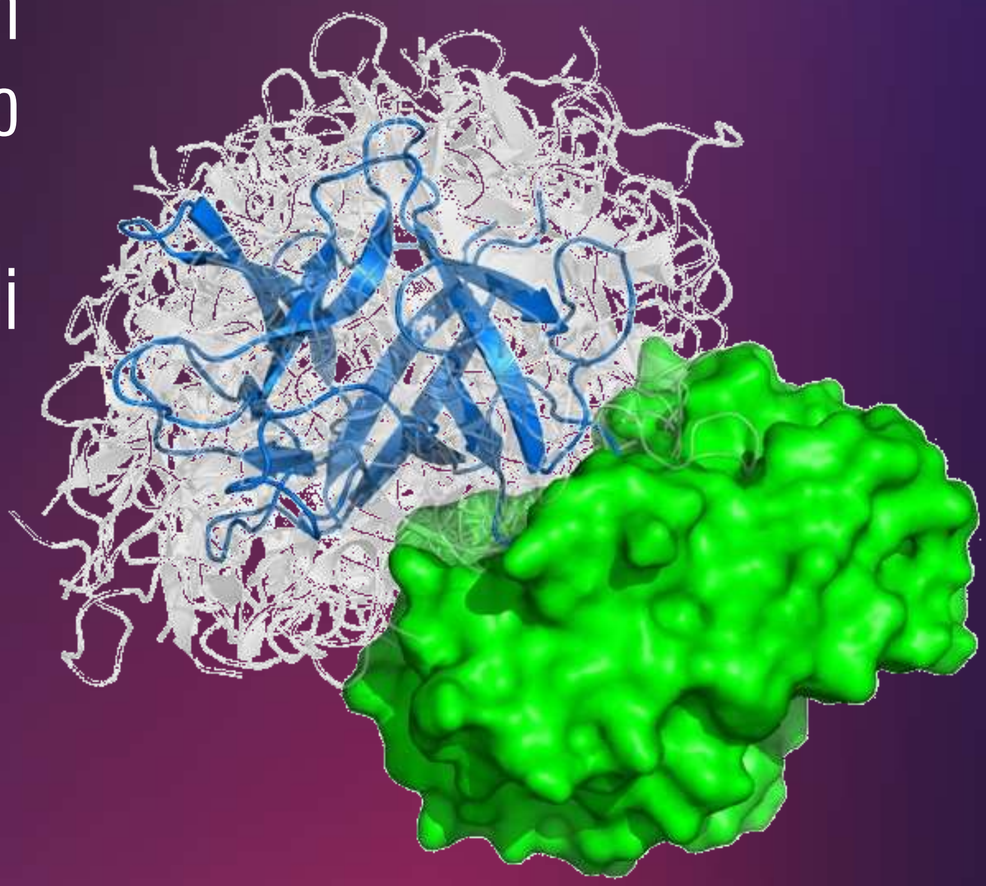
KNOWLEDGE PARTNER

FONDAZIONE Mondo digitale

Diritto alla conoscenza

INTRODUZIONE

- L’interazione tra compound e proteina è un task fondamentale per risparmiare tempo nella sperimentazione dei nuovi farmaci
- La IA può aiutare a predire la capacità di interazione tra farmaco e proteina!

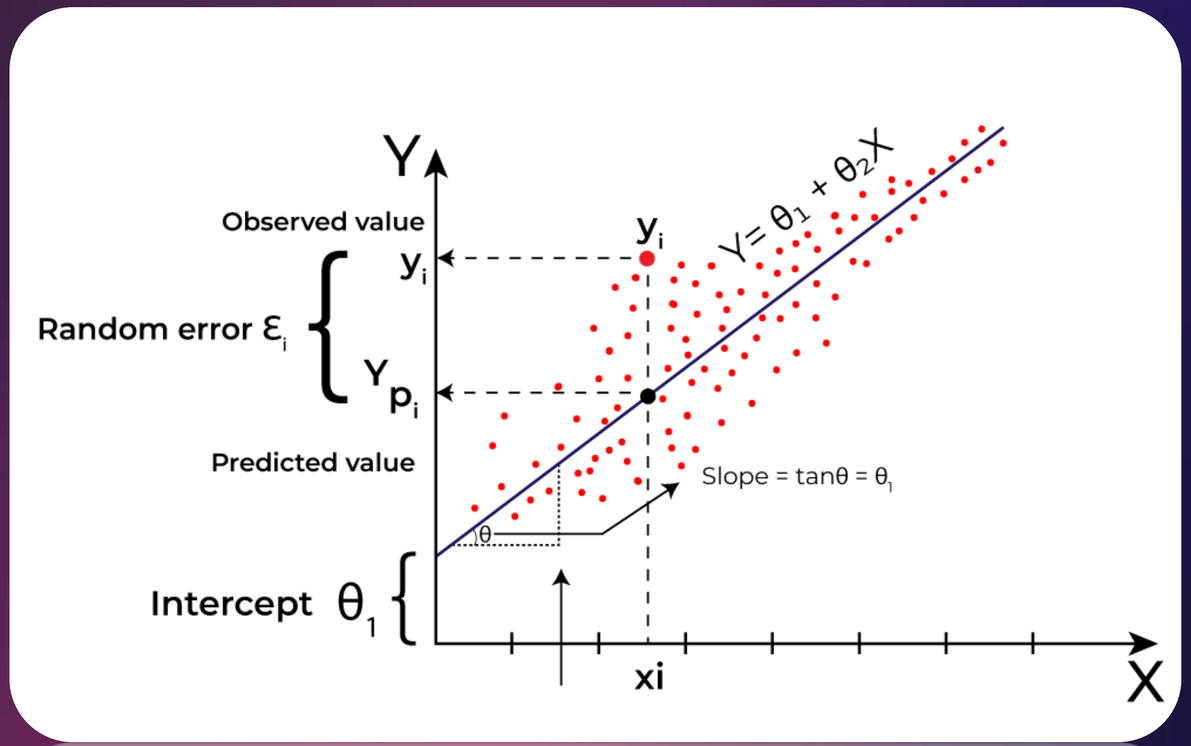


Per la challenge, abbiamo affrontato il problema di predire l’interazione tra proteina e farmaco usando la regressione lineare

IL MODELLO: REGRESSIONE LINEARE

- La regressione lineare è uno dei modelli statistici e di machine learning più semplici e utilizzati.
- Il suo obiettivo è stimare la relazione tra una variabile indipendente (o più variabili) e una variabile dipendente continua, attraverso una funzione lineare.
- In parole semplici: approssimiamo con una linea i dati da imparare!

$$y = b0 + b1x + \varepsilon$$



VANTAGGI E LIMITI DELLA REGRESSIONE LINEARE

✅ Vantaggi:

- Semplice da implementare
- Facilmente interpretabile
- Computazionalmente efficiente

❌ Limiti:

- Assume una relazione lineare tra le variabili
- Poco adatta per dati complessi o con relazioni non lineari
- Sensibile a valori anomali (outlier)

FLUSSO DI LAVORO

- **Lettura e pulizia dati (numerici)**
- Numero esempi nel training set: 1592
- Numero di esempi nel test set: 398

	molecular_weight	logp	h_bond_donors	h_bond_acceptors	rotatable_bonds	polar_surface_area	compound_clogp	protein_length	protein_pi	hydrophobicity	binding_site_size	mw_ratio
240	0.515256	0.383035	0.500	0.266667	0.428571	0.545892	0.197769	0.803544	0.627700	0.505971	0.607936	0.100888
809	0.532855	0.739252	0.250	0.533333	0.714286	0.450137	0.578315	0.194915	0.401999	0.621805	0.416374	0.338989
1458	0.485886	0.545640	0.250	0.466667	0.214286	0.537256	0.451557	0.337442	0.525487	0.285570	0.661332	0.211298
1910	0.373496	0.545640	0.125	0.333333	0.571429	0.492800	0.685520	0.989214	0.346101	0.515603	0.418492	0.052463
1949	0.535421	0.664195	0.250	0.266667	0.214286	0.650556	0.328335	0.357473	0.546418	0.311298	0.691991	0.223144
...	...	...	...	...	...	...	...	...	...	...	...	...
1130	0.549557	0.646862	0.625	0.266667	0.428571	0.566765	0.520014	0.881356	0.218518	0.564730	0.488519	0.098371
1294	0.352728	0.582578	0.500	0.533333	0.357143	0.337659	0.550388	0.675655	0.513084	0.579623	0.764325	0.077940
860	0.382438	0.514795	0.125	0.333333	0.285714	0.330367	0.425305	0.640986	0.580680	0.353154	0.597735	0.090834
1459	0.600465	0.503888	0.250	0.133333	0.428571	0.547366	0.571805	0.441448	0.348748	0.462891	0.648221	0.211543
1126	0.594510	0.365527	0.125	0.466667	0.642857	0.340976	0.529555	0.435285	0.633702	0.584017	0.449806	0.211843

FLUSSO DI LAVORO (2)

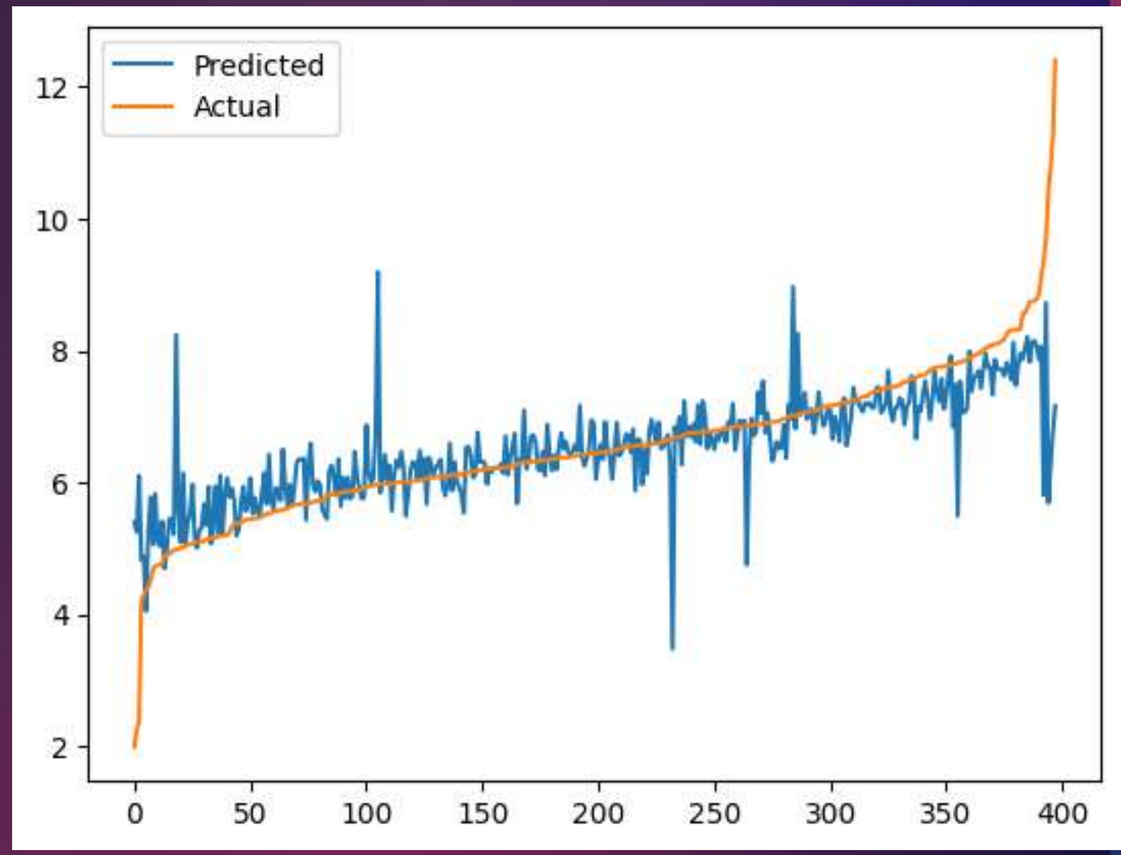
- Lettura e pulizia dati (numerici)
- Addestramento sul set di training
- Valutazione su test set (MSE)

L’obiettivo è predire il grado di interazione in modo più preciso possibile

RISULTATI

- E’ possibile predire l’interazione
- Possiamo sperimentare in laboratorio prima i compound che hanno un’interazione maggiore!

	compound_id	protein_id	Actual	Predicted
9	CID_01998	PID_125	unknown	7.584078
6	CID_01425	PID_125	unknown	7.311492
4	CID_00340	PID_125	unknown	7.099795
1	CID_00078	PID_125	unknown	6.910716
7	CID_01575	PID_125	unknown	6.894735
0	CID_00038	PID_125	unknown	6.436059
2	CID_00179	PID_125	unknown	6.417622
5	CID_00999	PID_125	unknown	6.365432
8	CID_01798	PID_125	unknown	6.037573
3	CID_00266	PID_125	unknown	5.752745



CORRIERE DELLA SERA

UNIVERSITÀ DI SIENA 1240

DIISM

CINECA

GRAZIE DELL'ATTENZIONE!

DISCLAIMER

ULTIME AVVERTENZE PRIMA DELLA RIVOLUZIONE

BIOMIND

INTELLIGENZA ARTIFICIALE PER L'INNOVAZIONE
IN FARMACIA E BIOINFORMATICA

KNOWLEDGE PARTNER

FONDAZIONE Mondo digitale
Diritto alla conoscenza