APPLICAZIONE DEI MODELLI PER L'INTELLIGENZA ARTIFICIALE SUI DATI DELLE IMPRESE BIOTECH IN ITALIA

RETE NEURALE E GRADIENT BOOSTING

Conferenza Nazionale di Statistica

VALENTINA PACCHIARELLI valentina.pacchiarelli@gmail.com | GAETANO COLETTA - Enea gaetano.coletta@enea.it | ROSSANA COTRONEO - Enea rossana.cotroneo@enea.it | ALESSANDRO ZINI - Enea alessandro.zini@enea.it

INTRODUZIONE

L'intelligenza artificiale consente di simulare i processi dell'intelligenza umana attraverso la creazione e l'applicazione di algoritmi integrati in un ambiente di calcolo dinamico. La loro capacità di catturare relazioni anche non lineari potenzialmente molto complesse favorisce un buon adattamento ai dati.
Per questo lavoro, sono stati utilizzati dei modelli di intelligenza artificiale per predire e integrare le mancate risposte parziali relative alla Rilevazione statistica sulle imprese nel campo delle biotecnologie consuntivo 2021 e proiezioni sul 2022 svolta dall'ENEA in collaborazione con Assobiotec e, successivamente, si procederà con le previsioni statistiche in avanti. L'analisi e l'elaborazione dei dati è stata effettuata tramite il software di programmazione visiva per l'analisi dei dati "Orange- data mining", il quale utilizza Python come linguaggio di programmazione.

OBIETTIVO

Stimare i dati mancanti per il dato del 2021, nonché la previsione per il dato del 2022, relativamente ad alcune delle variabili-chiave Rilevazione statistica sulle imprese nel campo delle biotecnologie consuntivo 2021 e proiezioni sul 2022 svolta dall'ENEA in collaborazione con Assobiotec

METODOLOGIA

Le variabili per le quali è stata effettuata la stima dei valori mancanti e poi la previsione sono: fatturato, spese per Ricerca e Sviluppo, spese per Ricerca e Sviluppo nel comparto biotech, numero degli addetti totali, numero degli addetti per Ricerca e Sviluppo quelli relativi al comparto biotech. Allo scopo è stata utilizzata la serie storica pregressa (2014-2022). Dopo aver svolto il processo di controllo e correzione dei dati sono stati testati due modelli di intelligenza artificiale, le Reti Neurali (NN) e il Gradient Boosting (GB). A corredo dell'applicazione dei modelli vengono presentate misure di accuratezza e precisione degli stessi.

Pre-processamento dati

Pulizia e correzione dei dati

Creazioni di sottoinsiemi di aziende in base al ciclo di vita dell'azienda stessa

Elaborazione dati

Creazione workflow per singola variabile e in base al ciclo di vita dell'azienda

Scelta dei modelli da utilizzare (NN e GB)

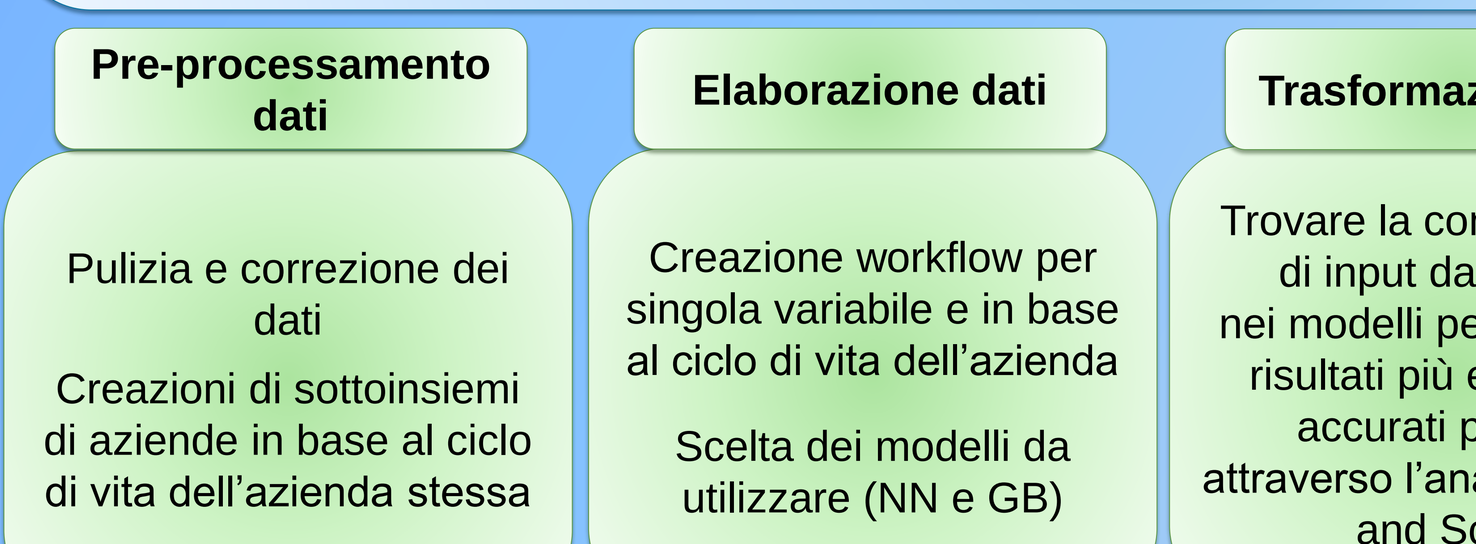
Trasformazione dati

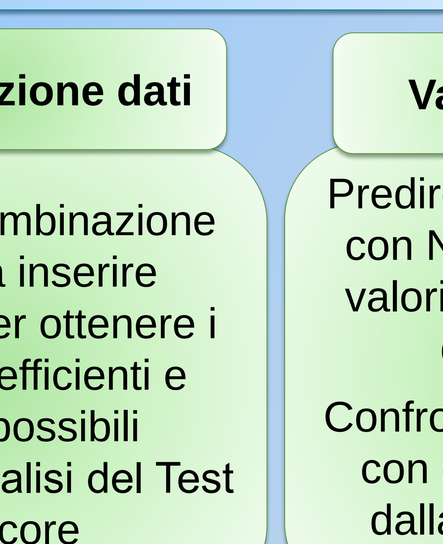
Trovare la combinazione di input da inserire nei modelli per ottenere i risultati più efficienti e accurati possibili attraverso l'analisi del Test and Score

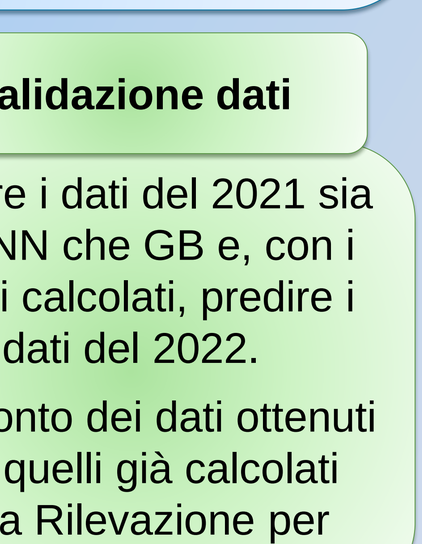
Validazione dati

Predire i dati del 2021 sia con NN che GB e, con i valori calcolati, predire i dati del 2022.

Confronto dei dati ottenuti con quelli già calcolati dalla Rilevazione per verificarne l'accuratezza







RISULTATI & CONCLUSIONI

Alla luce delle metriche maggiormente in uso, entrambi i modelli paiono in grado di fornire previsioni attendibili, con un soddisfacente livello di minimizzazione dell'errore in predizione.
Tra i due, il modello GB sembra esprimere una maggiore bontà di adattamento ai dati, calcolato in termini di R^2 ed errore quadratico medio, e stime più accurate e precise.

NN

FATTURATO	CORRELAZIONE	R^2
2021 CON NN	1,0000	1,0000
2022 CON NN	0,9986	1,0000
2021 CON GB	1,0000	1,0000
2022 CON GB	0,9986	0,9971

GB

RS INTRA BIOTECH	CORRELAZIONE	R^2
2021 CON NN	0,9993	0,9985
2022 CON NN	0,9909	0,9970
2021 CON GB	1,0000	0,9999
2022 CON GB	0,9911	0,9824

GB

ADDETTI RS	CORRELAZIONE	R^2
2021 CON NN	0,9999	0,9999
2022 CON NN	0,9992	0,9998
2021 CON GB	1,0000	1,0000
2022 CON GB	0,9993	0,9985

GB

SPESA RS	CORRELAZIONE	R^2
2021 CON NN	0,9512	0,9048
2022 CON NN	0,9390	0,8188
2021 CON GB	0,9980	0,9961
2022 CON GB	0,9396	0,8829

NN

ADDETTI TOTALI	CORRELAZIONE	R^2
2021 CON NN	1,0000	1,0000
2022 CON NN	0,9990	0,9999
2021 CON GB	1,0000	1,0000
2022 CON GB	0,9990	0,9980

NN

ADDETTI RS BIOTECH	CORRELAZIONE	R^2
2021 CON NN	0,9996	0,9991
2022 CON NN	0,9943	0,9983
2021 CON GB	0,9995	0,9990
2022 CON GB	0,9944	0,9887

GB

RS INTRA	CORRELAZIONE	R^2
2021 CON NN	0,9998	0,9995
2022 CON NN	0,9991	0,9991
2021 CON GB	1,0000	1,0000
2022 CON GB	0,9991	0,9982

NN

ADDETTI BIOTECH	CORRELAZIONE	R^2
2021 CON NN	0,9987	0,9973
2022 CON NN	0,7523	0,9947
2021 CON GB	1,0000	0,9999
2022 CON GB	0,7523	0,5660