Improving health and environmental statistics through artificial intelligence

L'AI PUO' RISOLVERE LE PROBLEMATICHE RELATIVE A DATI MANCANTI NELLE INDAGINI?

Conferenza Nazionale di Statistica

SIMONA CAFIERI - Istat cafieri@istat.it | **FRANCESCO ORTAME** - Istat ortame@istat.it | **FRANCESCO PUGLIESE** - Istat frpuglie@istat.it

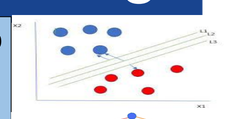
Background

Negli ultimi anni, gli Istituti nazionali di statistica sono sempre più chiamati a sviluppare quadri statistici su diversi argomenti, per contribuire a un processo decisionale informato. In particolare, in un'epoca di crescente interconnessione globale, è imperativo affrontare con forza le sfide emergenti che riguardano la sostenibilità ambientale, la salute e le disuguaglianze sociali. Anche in Italia questi temi sono di grande importanza in quanto tra gli elementi chiave degli indicatori BES (benessere equo e sostenibile) che sono alla base del documento di programmazione economica e finanziaria del governo :molti di questi indicatori sono calcolati a partire dai dati di indagine. In tali casi il problema principale è che dati incompleti o mancanti nei questionari possono compromettere l'accuratezza e l'affidabilità dei dati raccolti.

Lo scopo di questo lavoro è dunque quello di accertare se i rapidi progressi della tecnologia dell'informazione che hanno significativamente portato ad avanzamenti nel settore dell' intelligenza artificiale (IA), specialmente nei rami del Machine Learning (ML) e il Deep Learning (DL) possono essere applicati anche alle Statistiche Ufficiali per la risoluzione delle problematiche connesse a dati incompleti o mancanti.

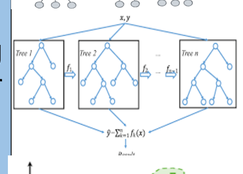
Metodi di machine e deep learning

Support Vector Machines (SVM): Un algoritmo che trova l'iperpiano ottimale per separare le diverse classi nello spazio delle caratteristiche.

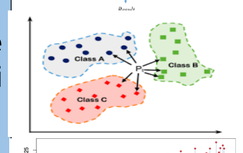


Decision Tree: Un modello non lineare che divide i dati in rami in base ai valori delle caratteristiche.

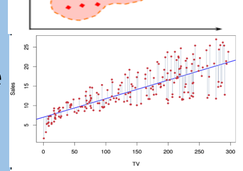
XGBoost (Extreme Gradient Boosting): Un algoritmo di boosting avanzato che combina le previsioni di diversi stimatori di base per migliorare le prestazioni del modello.



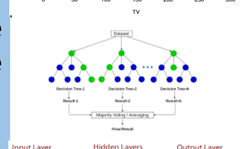
K-Nearest Neighbors (KNN): Un algoritmo basato sull'istanza che classifica i punti di dati in base alla classe maggioritaria tra i suoi vicini più prossimi.



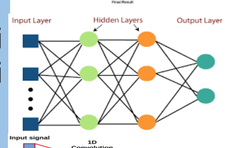
Regressione lineare: Un approccio lineare per modellare la relazione tra una var dipendente e una o più variabili indipendenti.



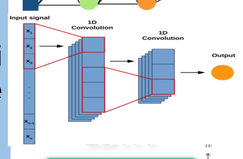
Random Forest: Un metodo di apprendimento collettivo che costruisce più alberi decisionali e fonde i loro risultati per migliorare l'accuratezza e la robustezza.



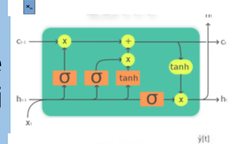
Multi-Layer Perceptron (MLP): Una classe di reti neurali artificiali feedforward con più strati di neuroni, utilizzate per compiti di classificazione e regressione.



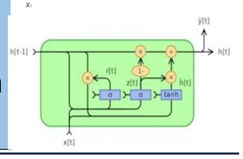
Convolutional 1D Neural Network: Variante delle reti neurali convoluzionali (CNN) specializzata per i dati unidimensionali, come le serie temporali o le sequenze.

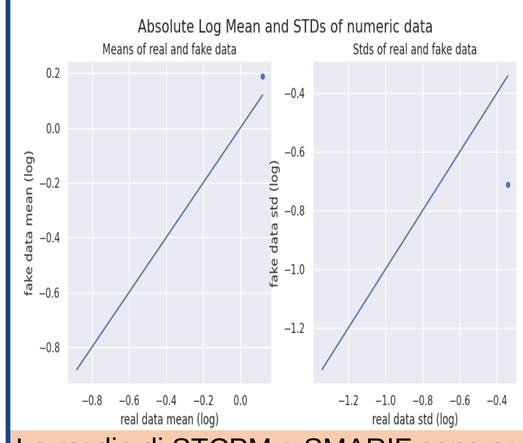


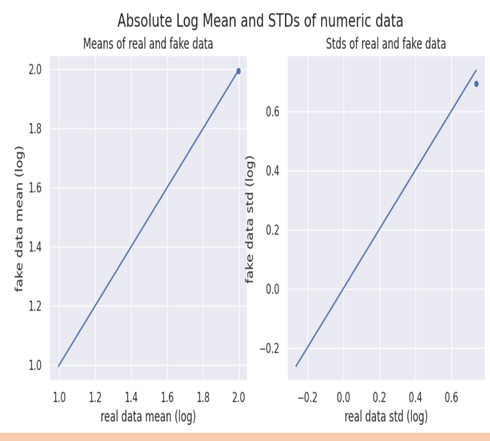
Long-Short Term Memory (LSTM): Un tipo di rete neurale ricorrente (RNN) progettata per catturare le dipendenze a lungo termine nei dati sequenziali utilizzando celle di memoria.



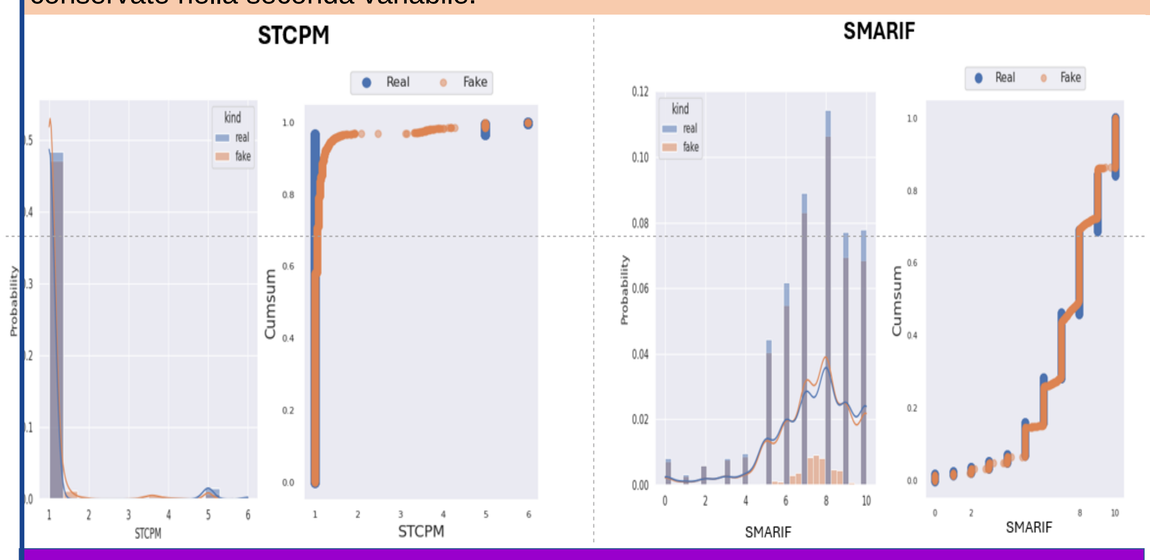
Gated Recurrent Unit (GRU): Un tipo di RNN simile alla LSTM, ma con un'architettura semplificata che la rende più veloce ed efficiente.



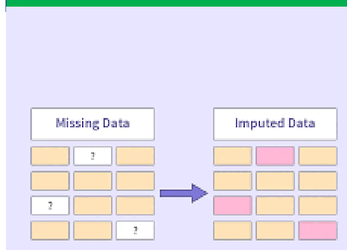




Le medie di STCPM e SMARIF sono correlate, mentre le varianze sono meglio conservate nella seconda variabile.



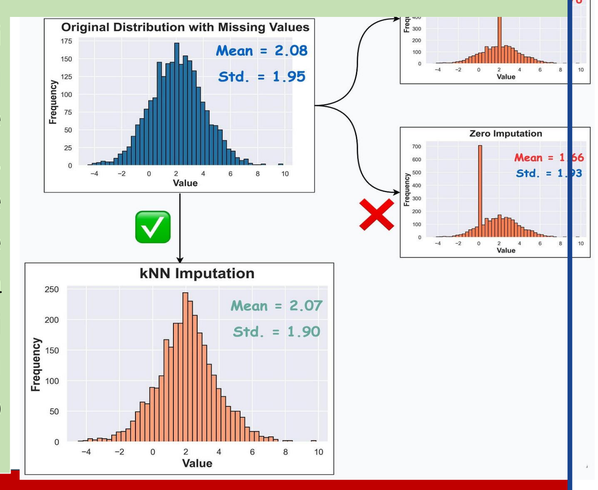
Metodologia



Una delle principali fonti di dati sociali e di salute delle famiglie in Italia è l'indagine Aspetti della vita quotidiana (AVQ), condotta ogni anno dall' ISTAT. In questa indagine dati mancanti nei questionari sono molto comuni e possono influenzare l'accuratezza e l'affidabilità dei risultati. Abbiamo quindi utilizzato questa rilevazione per fare un esperimento di imputazione con l'uso dell'Intelligenza Artificiale, utilizzando un dataset relativo all'indagine 2021, con 735 variabili

I valori mancanti nei dataset, spesso codificati come spazi vuoti o NaN, sono incompatibili con gli stimatori di scikit-learn, che richiedono che tutti i valori siano numerici e significativi. Un approccio di base è l'Analisi dei casi completi, in cui le righe (dropNA) o le colonne con valori mancanti (List-wise deletion) vengono scartate, Tuttavia, questo può portare a una

strategia efficace consiste nell'imputare i dati mancanti deducendoli dai dati disponibili Si possono utilizzare tecniche come le misure di tendenza centrale (media, mediana, ecc.). Tuttavia, è facile sottostimare o sovrastimare i valori reali, introducendo così dei bias nelle nostre stime, fenomeno che si verifica quando un algoritmo produce risultati sistematicamente distorti a causa di ipotesi errate, solitamente presenti nel set di dati o nel processo di apprendimento automatico.



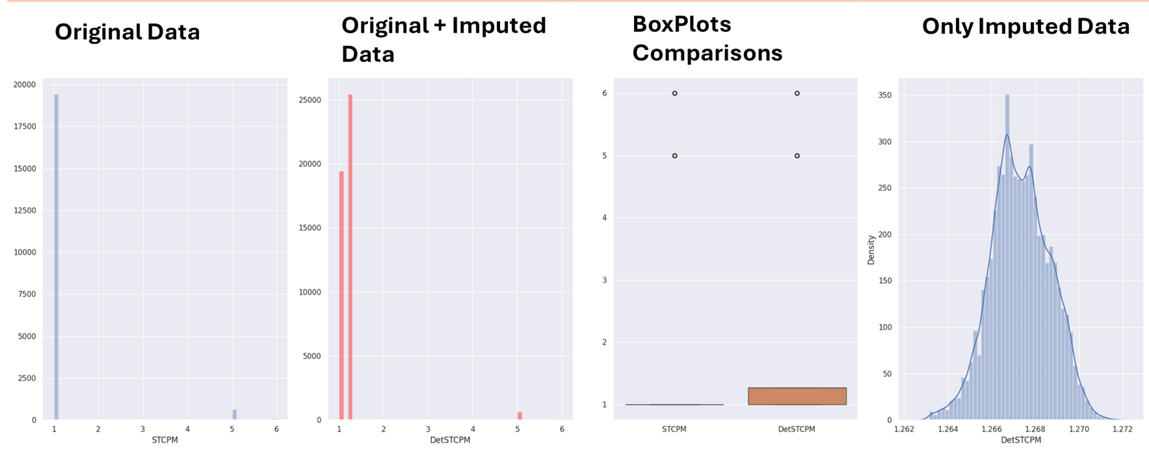
Risultati

Risultati dell'imputazione var. STCPM

Classifichiamo i modelli migliori in base all'errore quadratico medio (RMSE) in ordine decrescente, con il modello migliore che presenta l'errore più basso. Inoltre, consideriamo i modelli con il minor *overfitting* possibile, indicato da una bassa differenza tra l'RMSE del training set e quello del test set. In questo caso, LSTM risulta essere il

Model	Root Mean Squared Error on the Trainining Set	Root Mean Squared Error on the Test Set
Long-Short Term Memory	**0.6835**	**0.857**
Multi-Layer Perceptron	**0.6756**	**0.8625**
Convolutional 1D Neural Network	**0.6701**	**0.8581**
Support Vector Machines	**0.6703**	**0.8591**
Gated Recurrent Unit	**0.6722**	**0.8639**
K-Nearest Neighbours	**0.63**	**0.8676**
Logistic Regression	**0.61**	**0.8848**
Random Forest	**0.2569**	**0.8606**
XGBoostGradient Boosting	**0.2203**	**0.9025**
Decision Tree	**0**	**1.1208**

Come si può osservare, la distribuzione imputata è simile a quella originale (primi due grafici). L'unica differenza è nella varianza, come mostra l'intervallo interquartile nel terzo grafico. I dati imputati presentano una maggiore variabilità (ultimo grafico).

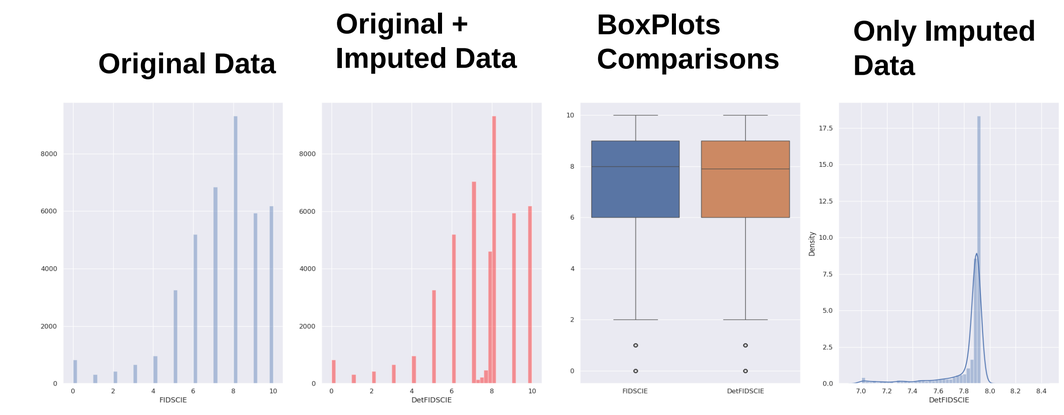


Risultati dell'imputazione var. SMARIF

L'addestramento di modelli di Machine Learning e Deep Learning rivela che le condizioni al contorno cambiano per ogni colonna e quindi il modello migliore è sempre diverso.

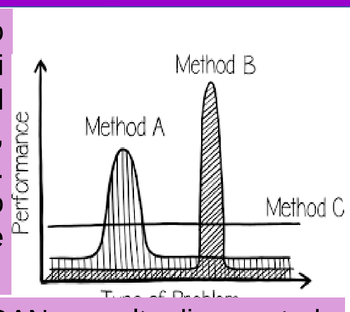
Model	Root Mean Squared Error on the Trainining Set	Root Mean Squared Error on the Test Set
Support Vector Machines	**2.1479**	**2.2083**
Gated Recurrent Unit	**2.091**	**2.157**
Multi-LayCer Perceptron	**2.1009**	**2.1689**
Convolutional 1D Neural Network	**2.1139**	**2.1825**
Long-Short Term Memory	**2.1064**	**2.1754**
K-Nearest Neighbours	**1.9205**	**2.1797**
Logistic Regression	**1.4796**	**1.7727**
XGBoost Gradient Boosting	**0.7809**	**1.6095**
Random Forest	**0.615**	**1.5987**
Decision Tree	**0**	**2.2062**

In questo caso, i modelli SVM sono i migliori. Entrambe le distribuzioni sono quas identicheed anche i boxplot e non sono influenzate dal processo di imputazione.



Conclusioni e Bibliografia

Secondo il Teorema del No-Free-Lunch, non esiste un unico modello di apprendimento automatico migliore per ogni colonna. La forza di questo metodo sta nel trovare sempre il modello migliore per ogni variabile da imputare, accuratamente selezionato da un'ampia gamma di modelli. Un'analisi approfondita ci permette di scegliere il modello migliore, di imputare i dati mancanti con esso e di valutare successivamente la qualità dei dati imputati.



Sviluppi futuri :utilizzo di altri modelli, come Transformer e GAN, e scelta di un metodo di imputazione a regressione stocastica che preservi meglio la varianza tra la distribuzione originale e quella imputata.

Bibliografia:

Pinar Cihan Deep learning-based approach for missing data imputation , Eskişehir Teknik Üniversitesi Bilim 2020

Ballard DH. Modular Learning in Neural Networks. In: AAAI, 1987. pp 279-284

Sebastian Raschka, Vahid Mirjalili Python Machine Learning-PACKT Birmingham

ISTAT, Indagine Aspetti della vita quotidiana 2021