Borse di studio GARR
Orio Carlini

L'uso del Machine Learning nella previsione probabilistica degli eventi vulcanici dell'Etna

Giulio Domina

Istituto Nazionale di Geofisica e Vulcanologia - Osservatorio Etneo

Consortium GARR
THE ITALIAN EDUCATION & RESEARCH NETWORK

Introduzione

Sviluppare un sistema di ***early detection probabilistico*** per prevedere eruzioni dell'Etna con precisione e tempestività, utilizzando dati sismici, geochimici e geofisici.

Analizzare grandi volumi di dati multivariati per identificare schemi che precedono le eruzioni vulcaniche.

FASI: Pipeline del progetto

1. Revisione del contesto esistente
2. Ingegneria dei dati
3. Creazione e valutazione dei modelli

prolessi

Attività Parossistica

Attività Stromboliana

Primo Trimestre :

Analisi del contesto vulcanico Etneo

- Analisi della letteratura sui fenomeni vulcanici e sulle applicazione del machine learning
- Creazione del catalogo di parossismi dal 2011 al 2022 (116 eventi parossistici)
- Definizione delle metriche e benchmark di valutazione
- Richiesta accesso alle reti permanenti

Primo Trimestre :

Benchmark

True Positive Rate (TPR): calcolata come la frazione degli eventi target correttamente identificati dall'allerta. Un valore più vicino all'uno indica una maggiore capacità del sistema di identificare correttamente gli eventi reali.

$$Probability\ of\ detection\ P(Allerta|evento) = \#\frac{evento\ \&\ Allerta}{\#eventi} = X$$

False Discovery Rate (FDR): calcolata come la percentuale di allerte che non corrispondono ad un evento target reale. Un valore più basso indica una minor frequenza di falsi allarmi.

$$Probability\ of\ false\ alarm\ P(no.evento|Allerta) = \#\frac{Allerta\ \&\ no.evento}{\#Allerta} = X$$

Tempo medio di anticipo dell'allerta (Ta): Indica il tempo medio di anticipo con cui l'allerta è stata rilevata rispetto all'inizio effettivo dell'evento. Un valore positivo indica un anticipo, mentre un valore negativo indica un ritardo.

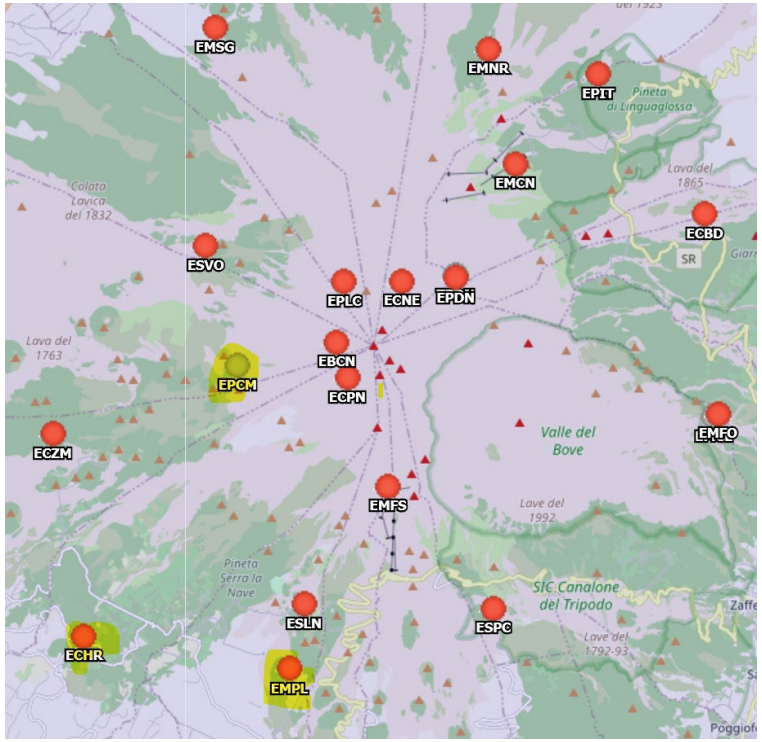
Frazione di tempo in allerta (FTA): Rappresenta la percentuale del tempo totale di funzionamento del sistema in cui l'allerta è attiva.

Secondo Trimestre

Raccolta dati, catalogo eventi e confronto modelli deep learning

- 16 stazioni sismiche dal 2011 al 2022

  feature: *RMS* (campionamento a 10 min)



Secondo Trimestre

Elaborazione dati

Riempimento dei dati mancanti:

< ora (**interpolazione lineare**)

> ora

- **Imputazione Iterativa con Valore Minimo Zero**
- Imputazione Semplice con Media
- Imputazione Semplice con Mediana
- Imputazione Semplice con Valore più Frequente
- Imputazione Semplice con Valore Costante

Outlier:

- **Metodo della Deviazione Standard**
- Metodo dell'Intervallo Interquartile (IQR)
- Z-score
- Distanza Euclidea
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Secondo Trimestre

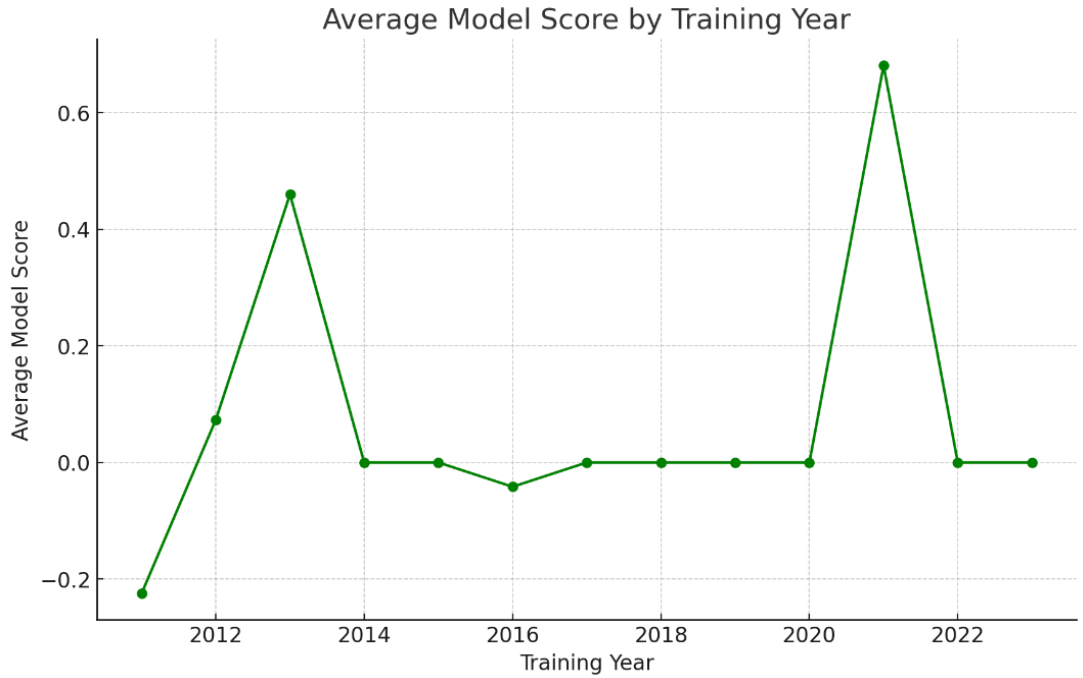
Modelli scelti per l'analisi di dati sequenziali

1. **Dense Neural Network (Dense)**: Strati completamente connessi, diretto ma meno efficace per relazioni temporali profonde.
2. **Gated Recurrent Unit (GRU)**: Variante leggera delle reti ricorrenti per dipendenze temporali nei dati sequenziali.
3. **Long Short-Term Memory (LSTM)**: Mantiene informazioni su lunghi intervalli di tempo, ideale per pattern prolungati nei segnali sismici.
4. **Pooling Convolutional Networks (1D, 2D, 3D)**:
    1. **1D**: Pooling lungo la dimensione temporale.
    2. **2D**: Pooling spaziale e temporale.
    3. **3D**: Relazioni complesse nei dati volumetrici.
5. **Transformer**: Efficace per dipendenze a lungo raggio; elabora informazioni in parallelo rispetto alla sequenzialità delle reti ricorrenti.

Secondo Trimestre

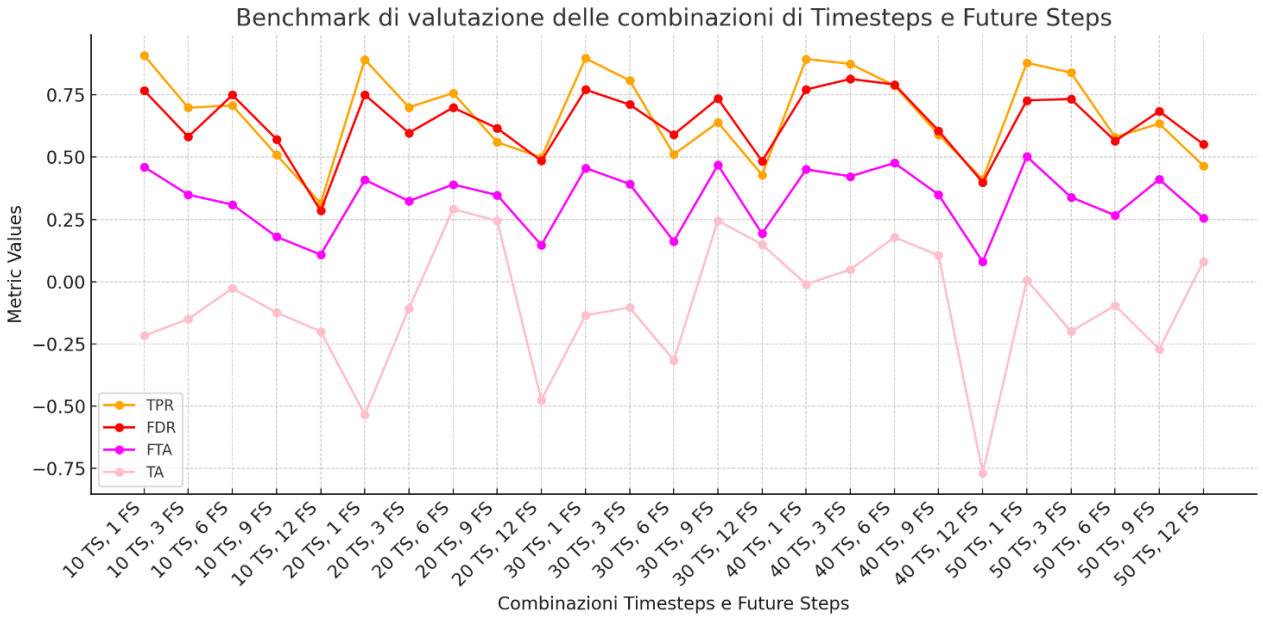
Fase di tuning (anno di allenamento 2021)

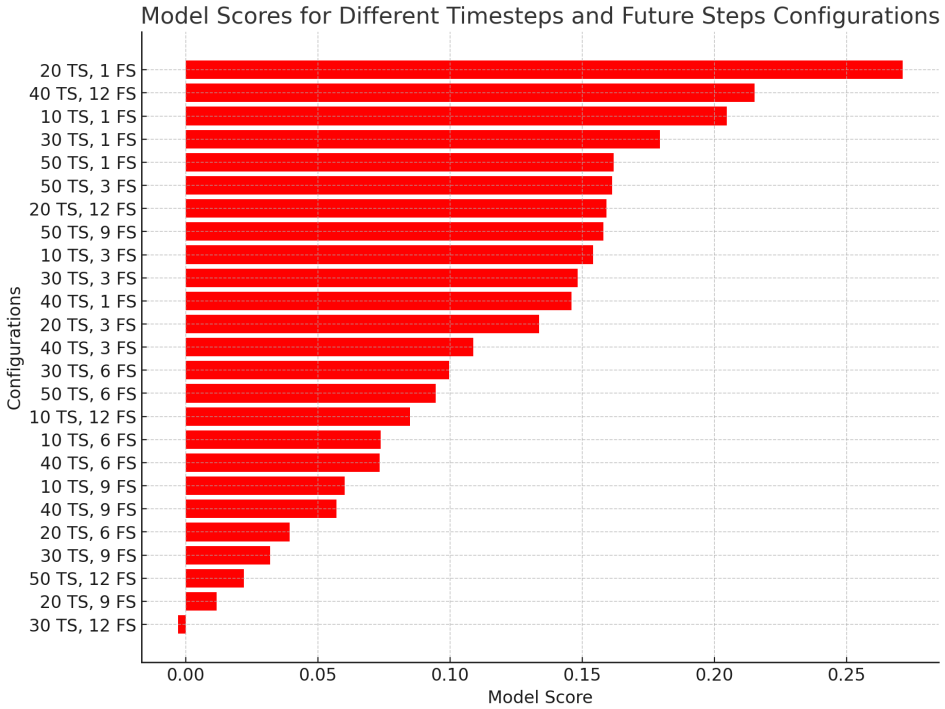
$$Model_{score} = \frac{1}{4} * TPR + \frac{1}{4} * FTA - \frac{1}{4} * FDR - \frac{1}{4} * TA$$



Secondo Trimestre

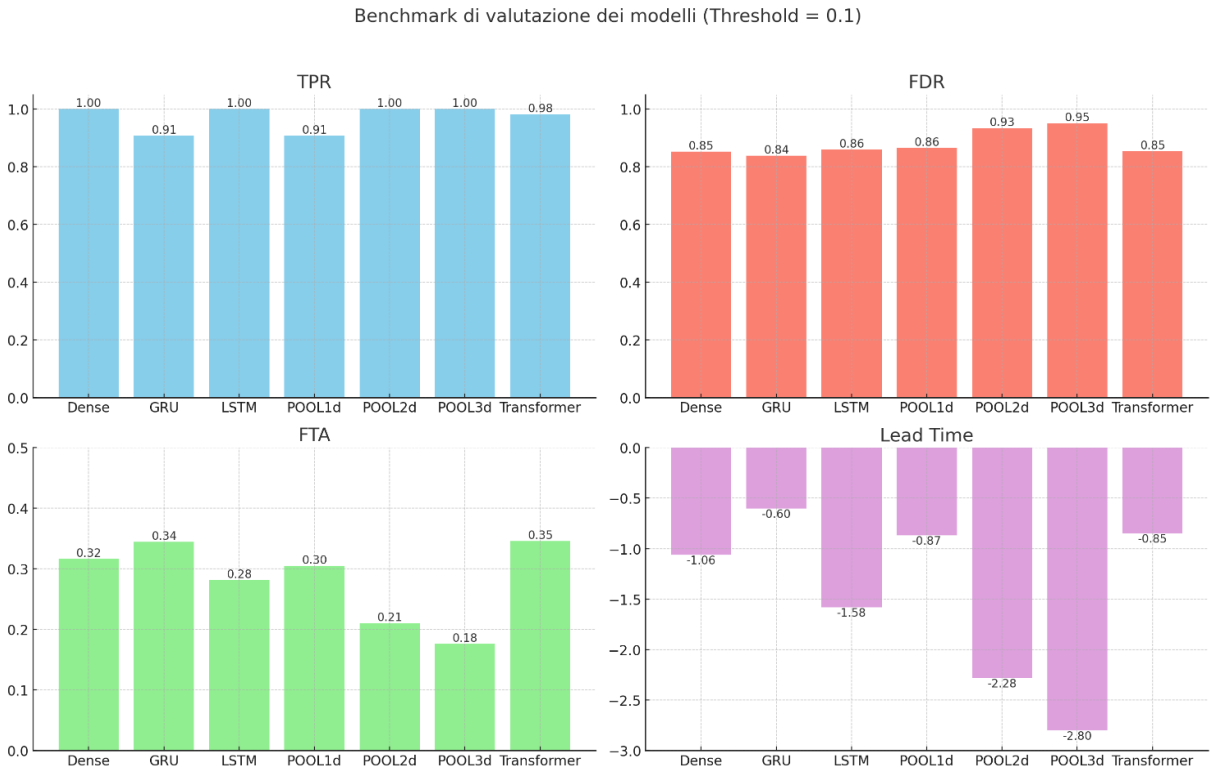
Fase di tuning (time_step/future_step)

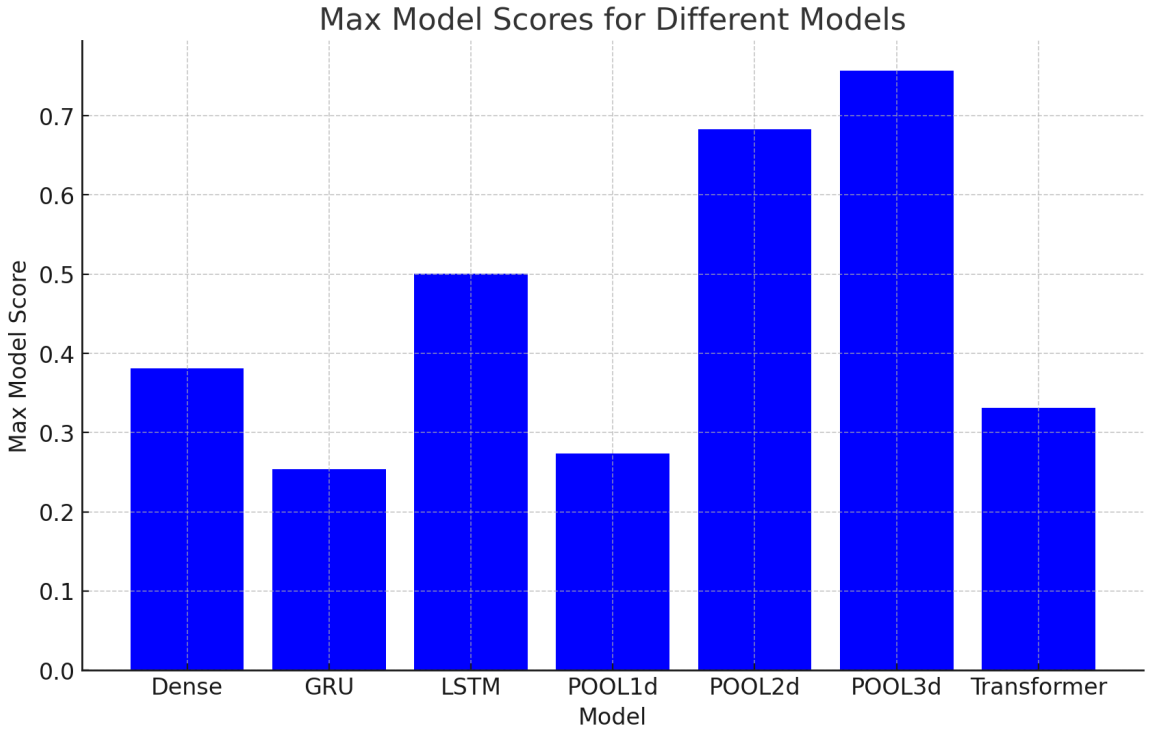




Secondo Trimestre

Benchmark modelli

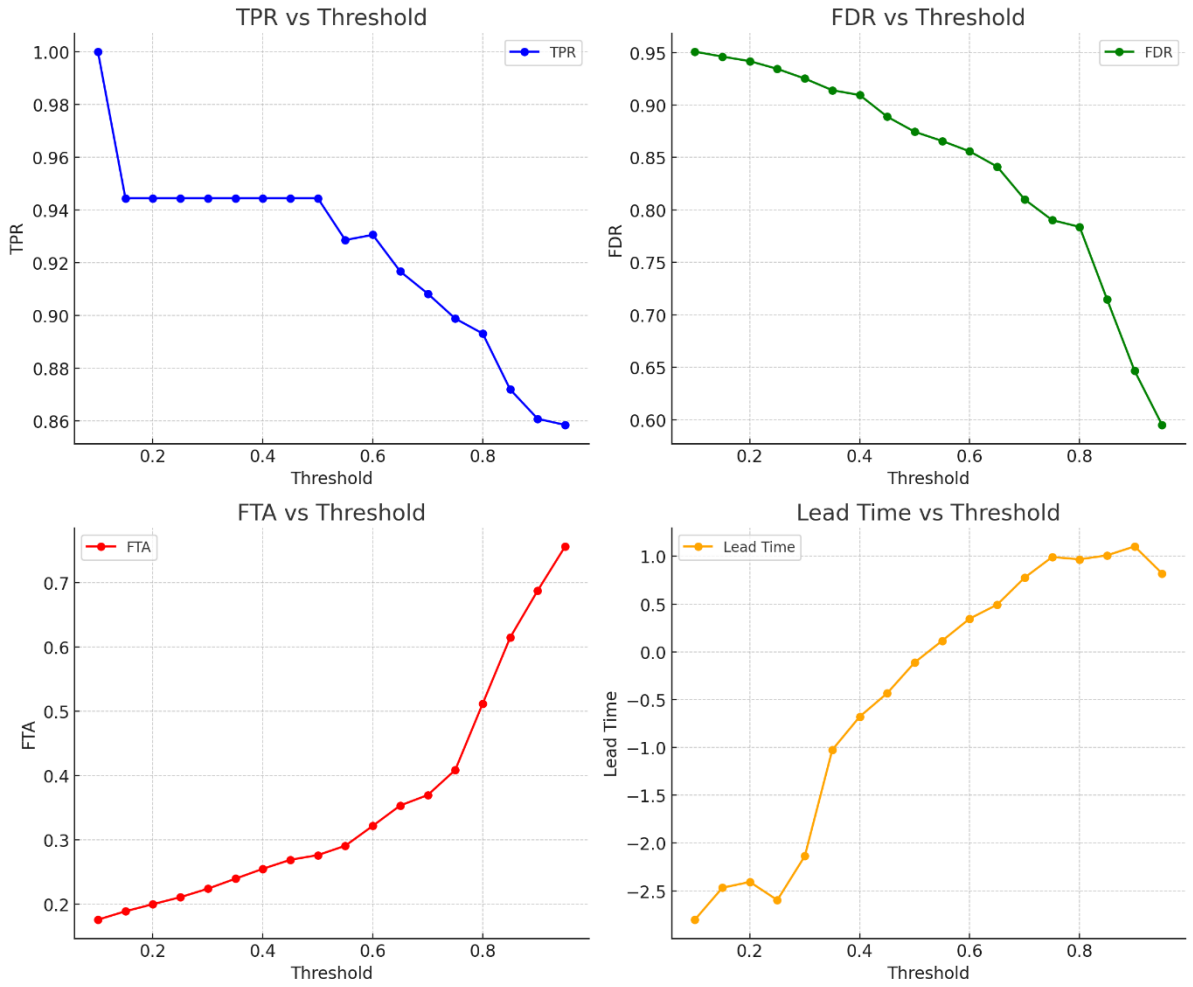




Secondo Trimestre

Miglior modello Pool_3D

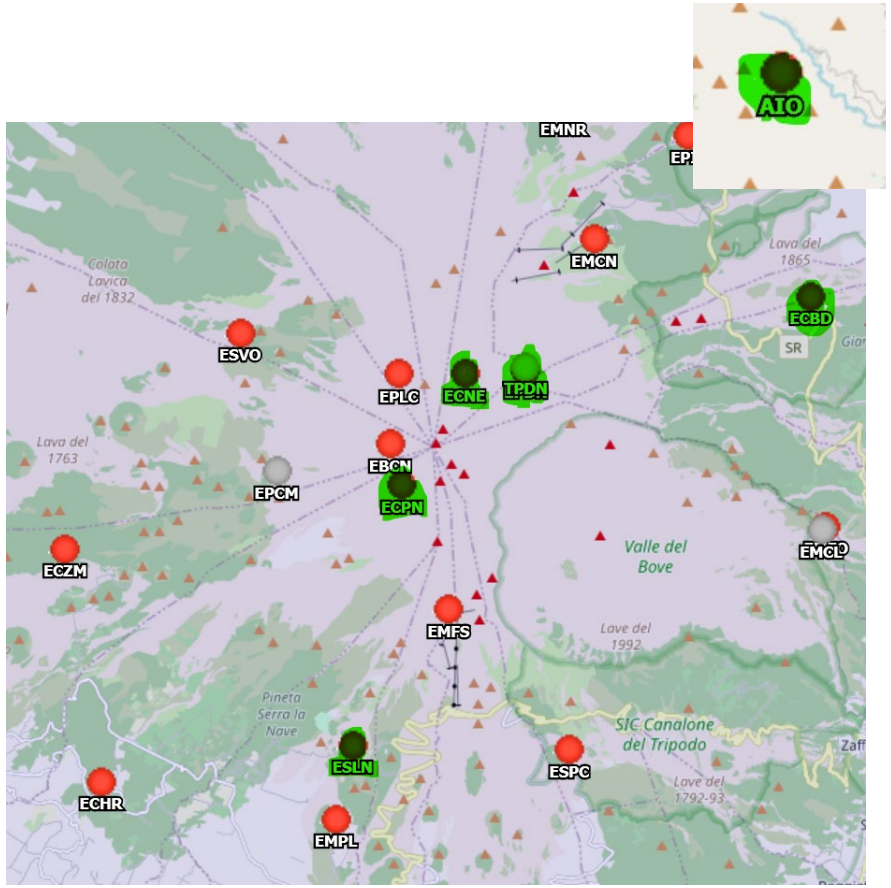
Il **threshold** rappresenta un valore di soglia utilizzato per classificare o decidere tra due stati, come eventi positivi o negativi. Modificando il threshold, si influenzano metriche permettendo di bilanciare la sensibilità e la specificità del modello o sistema.



Terzo Trimestre

Raccolta Dati grezzi e gestione dati 2021

- Dati sismici -100hz- (5 stazioni)
- Dati infrasuono -50hz- (4 stazioni)
- Dati clinomentrici -campionamento 1min- (4 stazioni)
  - Asse Y
  - Asse X



Terzo Trimestre

Estrazione feature sismico

- Dati sismici 3 componenti (Z,N,E)

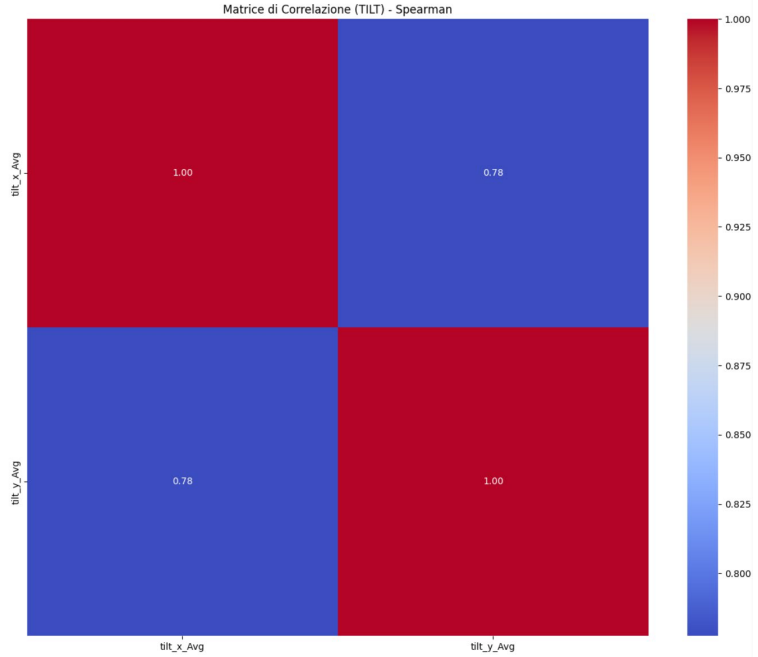
1. Energy
2. RMS
3. Max Amplitude
4. Power Band Features
5. Spectral Entropy
6. Standard Deviation
7. Variance
8. Skewness
9. Kurtosis
10. Crest Factor
11. Wavelet Coefficients
12. Peak Count
13. Horizontal to Vertical Ratio
14. Motion Direction
15. Rolling SNR

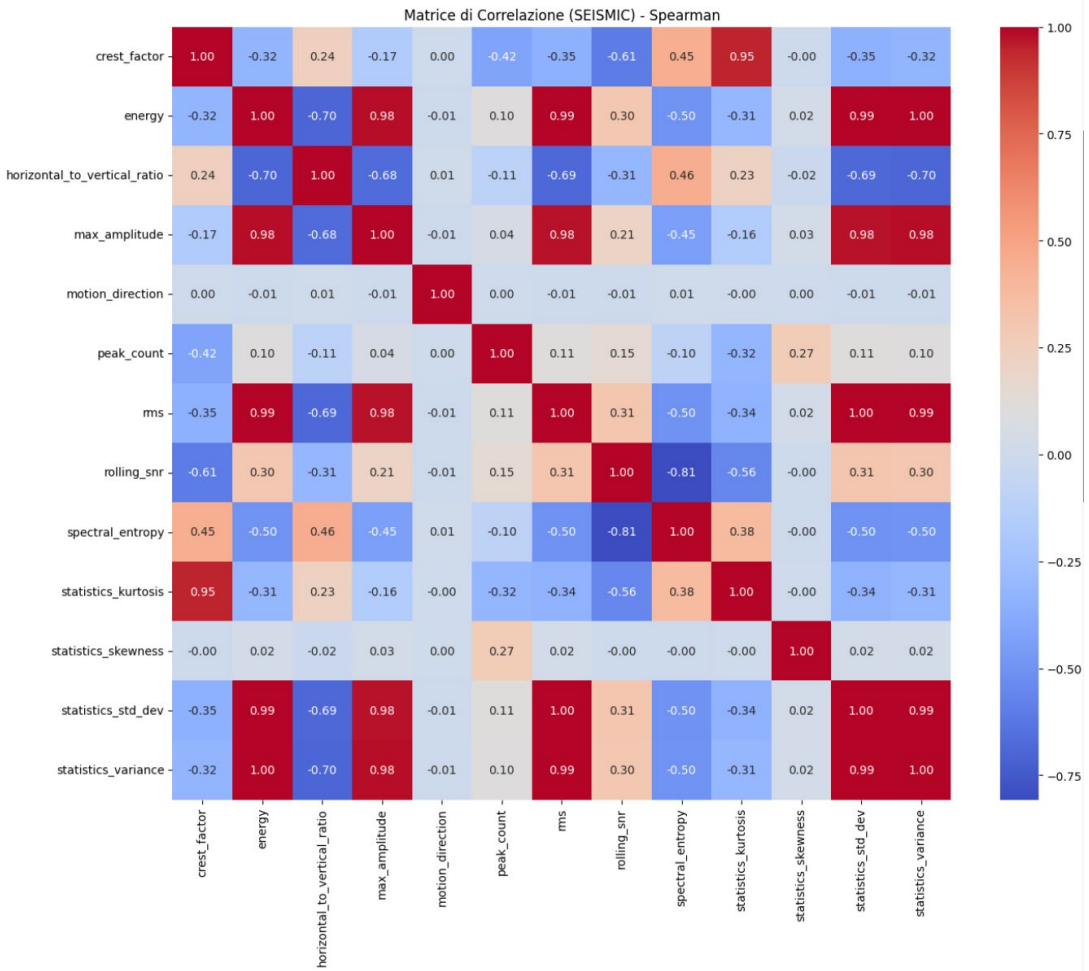
Dati mancanti: imputazione per media mobile ponderata (dalla distanza geografica delle stazioni)

Outlier: Metodo della Deviazione Standard

Terzo Trimestre

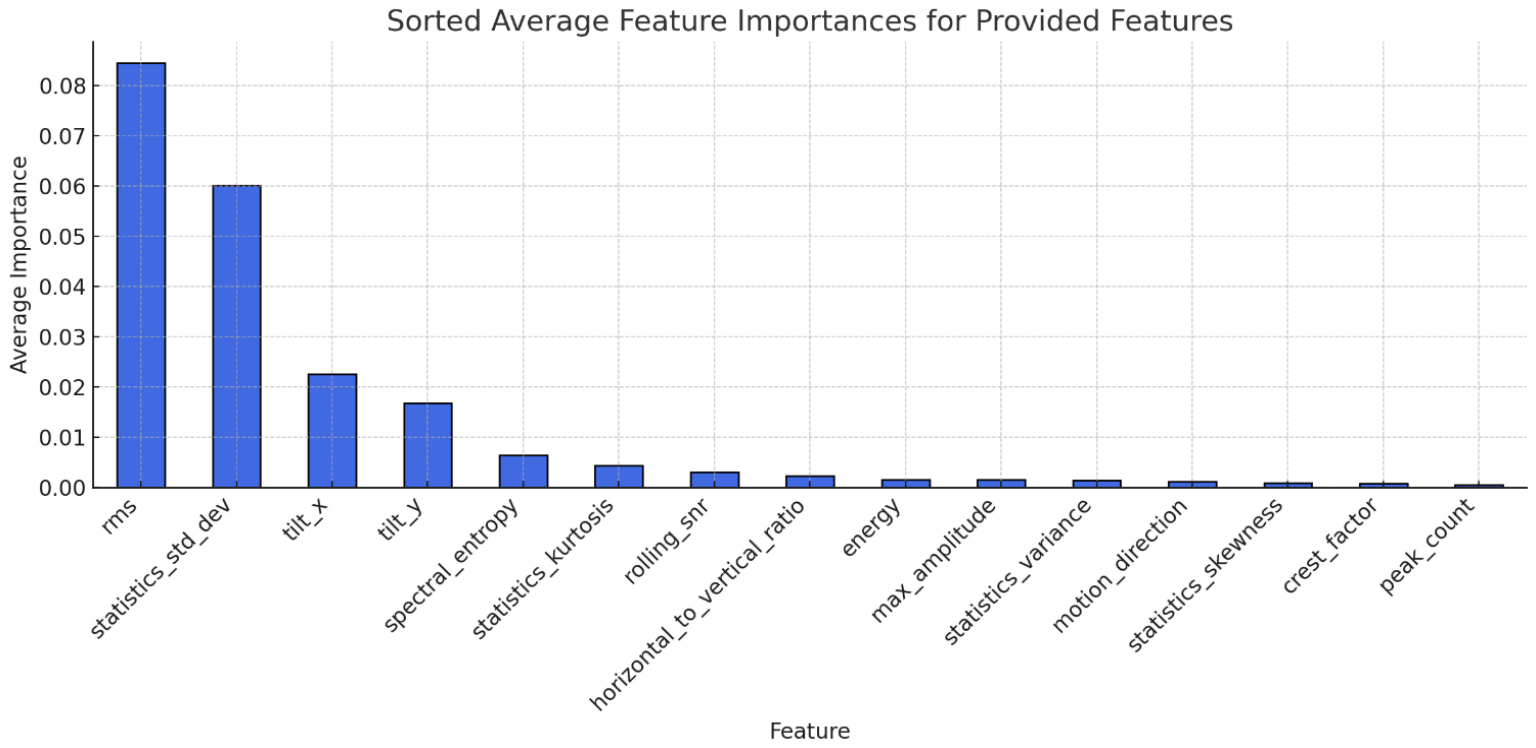
Matrice di correlazione (Spearman) tra dati clinometrici e sismici





Terzo Trimestre

Feature importance sismico e tilt



Quarto Trimestre

Estrazione feature infrasuono

- Dati infrasuono componente ‘O’

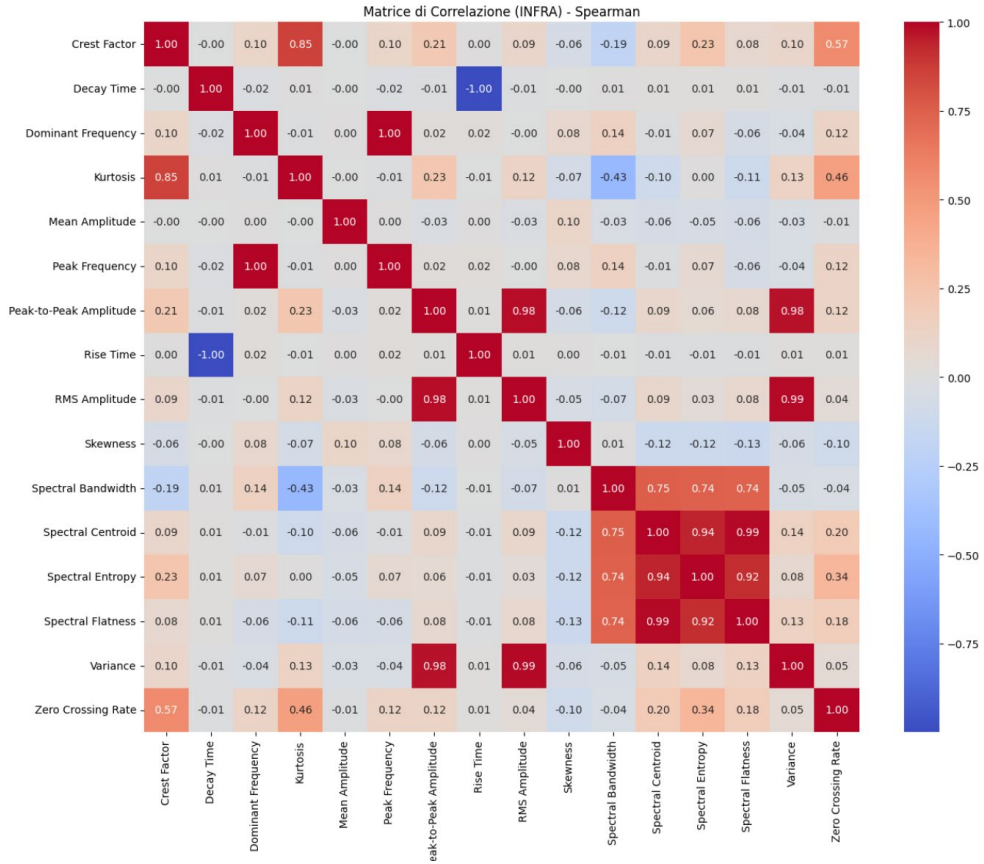
1. Peak-to-Peak Amplitude
2. RMS Amplitude
3. Zero Crossing Rate
4. Rise Time
5. Decay Time
6. Mean Amplitude
7. Variance
8. Skewness
9. Kurtosis
10. Crest Factor
11. Rolling SNR
12. Distortion
13. Dominant Frequency
14. Spectral Centroid
15. Spectral Bandwidth
16. Spectral Entropy
17. Spectral Flatness
18. Peak Frequency

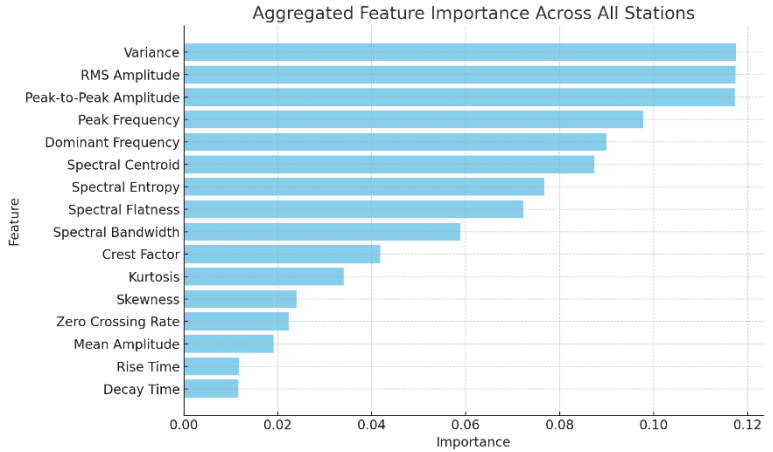
Dati mancanti: imputazione per media mobile ponderata (dalla distanza geografica delle stazioni)

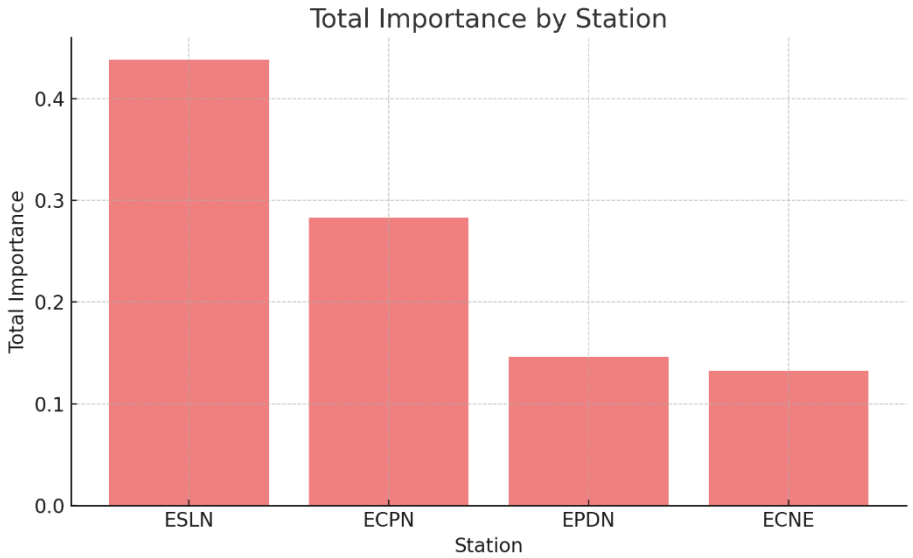
Outlier: Metodo della Deviazione Standard

Quarto Trimestre

Matrice di confusione (Spearman) e feature importance







Quarto Trimestre

Selezione e creazione modello per la feature selection

1. Addestrare un modello su tutte le feature e valutare le performance.

2. Disattiva una feature alla volta e ricalcola la performance.

3. Identifica le feature che causano un calo significativo di performance.

4. Selezione delle feature per l'allenamento dei modelli finali

Creazione e comparazione delle performance dei modelli finali

Lettera Aperta

Promuovere l'Open Science per una Ricerca più Trasparente e Inclusiva

Introduzione

- La condivisione di conoscenze e risorse è una pratica etica e strategica.
- L'attuale cultura di scarsa condivisione frena l'efficienza e la collaborazione.
- Open Science e Open Data amplificano l'impatto della ricerca scientifica.

Normative Rilevanti sull'Open Data:

1. *Direttiva (UE) 2019/1024*: Promuove il riutilizzo dei dati pubblici.
2. *Decreto Legislativo n. 200/2021*: Recepimento della direttiva UE in Italia.
3. *Linee Guida AgID sugli Open Data*: Standard tecnici per dati aperti nella PA.

Lettera Aperta

Promuoere l'Open Science per una Ricerca più Trasparente e Inclusiva

Proposte Concrete

Piattaforma Open Data INGV: Centralizzazione e accessibilità dei dati.

Policy Istituzionale: Linee guida chiare per la condivisione.

Formazione Interna: Workshop sui benefici e le modalità dell'open science.

Collaborazioni Esterne: Partnership con enti impegnati nell'open science.

Incentivi e Riconoscimenti: Valorizzazione delle buone pratiche.

Un Invito alla Comunità INGV

Abbracciare l'Open Science significa intraprendere un percorso di crescita collettiva che unisce etica e strategia, ponendo l'INGV come modello di eccellenza e innovazione nella comunità scientifica. Attraverso il dialogo e l'azione condivisa possiamo costruire un futuro in cui la trasparenza e la collaborazione siano al centro del nostro operato.

Borse di studio GARR
Orio Carlini

Grazie per l'ascolto

Q&A

Consortium GARR
THE ITALIAN EDUCATION & RESEARCH NETWORK

Bibliografia

- Aloisi, M., Bonaccorso, A., Cannavò, F., Currenti, G., & Gambino, S. (2020). The 24 December 2018 eruptive intrusion at Etna volcano as revealed by multidisciplinary Continuous deformation Networks (CGPS, borehole strainmeters and tiltmeters). *Journal of Geophysical Research. Solid Earth*, *125*(8). https://doi.org/10.1029/2019jb019117
- Bonaccorso, A., Campisi, O., Falzone, G., & Gambino, S. (2004). Mt. Etna: Volcano Laboratory. In *Geophysical monograph*. https://doi.org/10.1029/gm143
- Calvari, S., & Nunnari, G. (2022). Comparison between Automated and Manual Detection of Lava Fountains from Fixed Monitoring Thermal Cameras at Etna Volcano, Italy. *Remote Sensing*, *14*(10), 2392. https://doi.org/10.3390/rs14102392
- Cannavò, F., Cannata, A., Cassisi, C., Di Grazia, G., Montalto, P., Prestifilippo, M., Privitera, E., Coltelli, M., & Gambino, S. (2017). A multivariate probabilistic graphical model for real-time volcano monitoring on Mount Etna. *Journal of Geophysical Research. Solid Earth*, *122*(5), 3480–3496. https://doi.org/10.1002/2016jb013512
- Carniel, Roberto, and Silvina Guzman. "Machine learning in volcanology: a review."
- *Volcanoes-Updates in Volcanology* (2021).
- Kapoor, Amita, et al. *Deep Learning with TensorFlow and Keras: Build and deploy supervised, unsupervised, deep, and reinforcement learning models*. Packt Publishing Ltd, 2022.
- Nelson, Hala. *Essential Math for AI*. " O'Reilly Media, Inc.", 2023.
- Papale, P. (2018). Global time-size distribution of volcanic eruptions on Earth.
- *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-25286-y
- Use of machine learning to analyze and – hopefully – predict volcano activity. (2017).
- *Acta Polytechnica Hungarica*, *14*(3). https://doi.org/10.12700/aph.14.3.2017.3.12
- Salerno, G. G., M. R. Burton, C. Oppenheimer, T. Caltabiano, V. I. Tsanev, and N. Bruno (2009), Novel retrieval of volcanic SO2 abundance from ultraviolet spectra, J. Volcanol. Geotherm. Res., 181, 141–153, doi:10.1016/j.jvolgeores.2009.01.009
- Li, Xin, et al. "Big Data in Earth system science and progress towards a digital twin." *Nature Reviews Earth & Environment* 4.5 (2023): 319-332.
- Foto di Casadipagliafelcerossa