UNIVERSITA' DEGLI STUDI DI CATANIA

Master's Degree in Data Science

Department of Mathematics and Computer Science(LM-Data)

Multiparametric analysis in Volcanic Environment

Final Thesis

Author:
Gaetano De Angelis

Supervisor:
Prof. Sebastiano Battiato

INGV CO-SUPERVISOR:
Dr. Salerno Giuseppe
Dr. Reitano Danilo

Academiic Year 2024-2025

1. Introduction

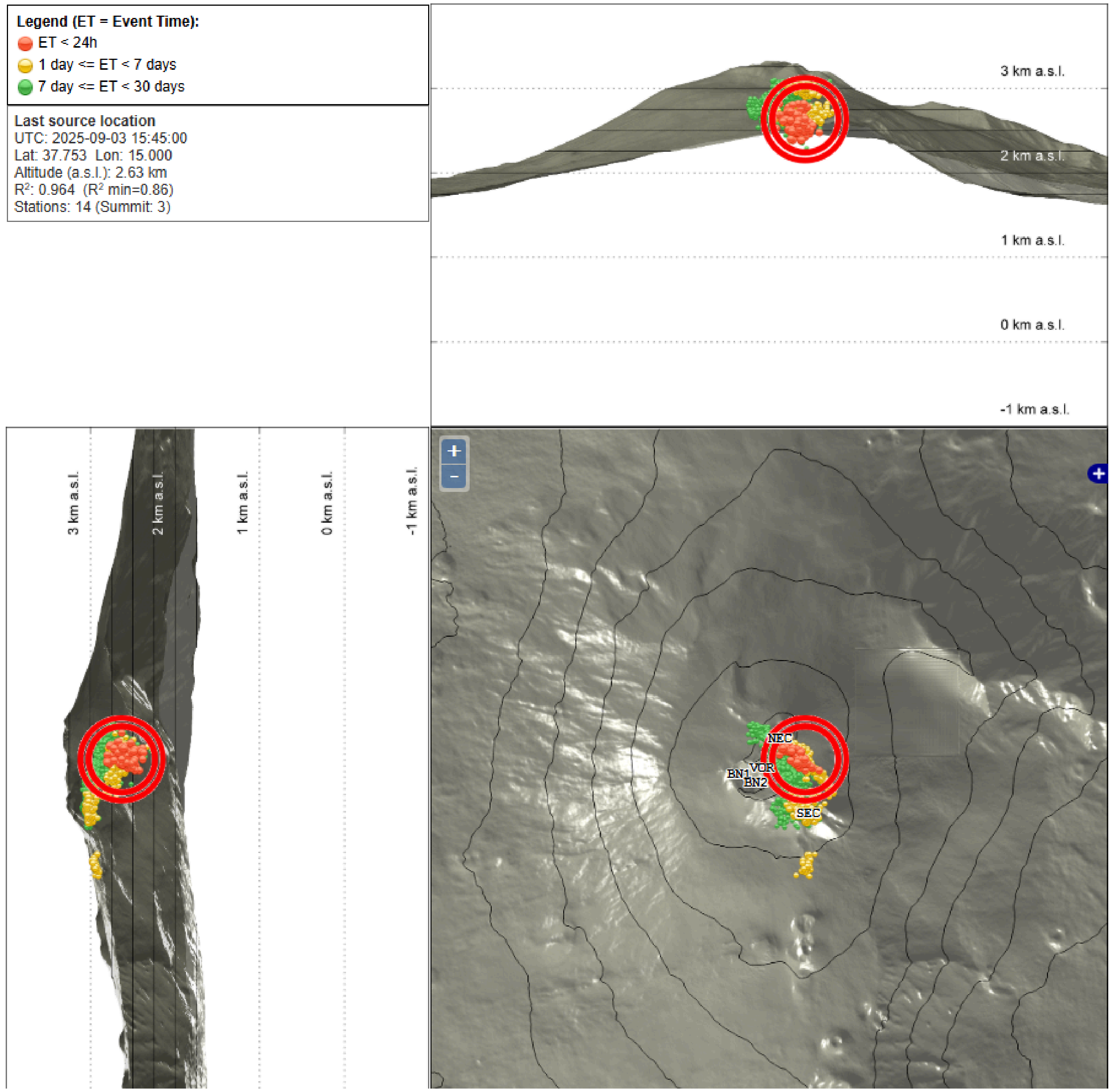
Data drive decisions, improve efficiency, and boost performance. In natural hazard monitoring, timely and accurate data are crucial for safety.

- **Volcanic monitoring**
  - INGV manages thousands of data daily via sensor networks..
- **Case study: Etna 2018**
  - Focus on eruption affecting Zafferana Etnea in 2018.
- **Data analysis process**
  - Managing data and Correlation analysis between variables and features..
- **Model implementation**
  - Machine and Deep Learning simulations to predict earthquakes within 24 hours .

2. Something about volcanic aspects

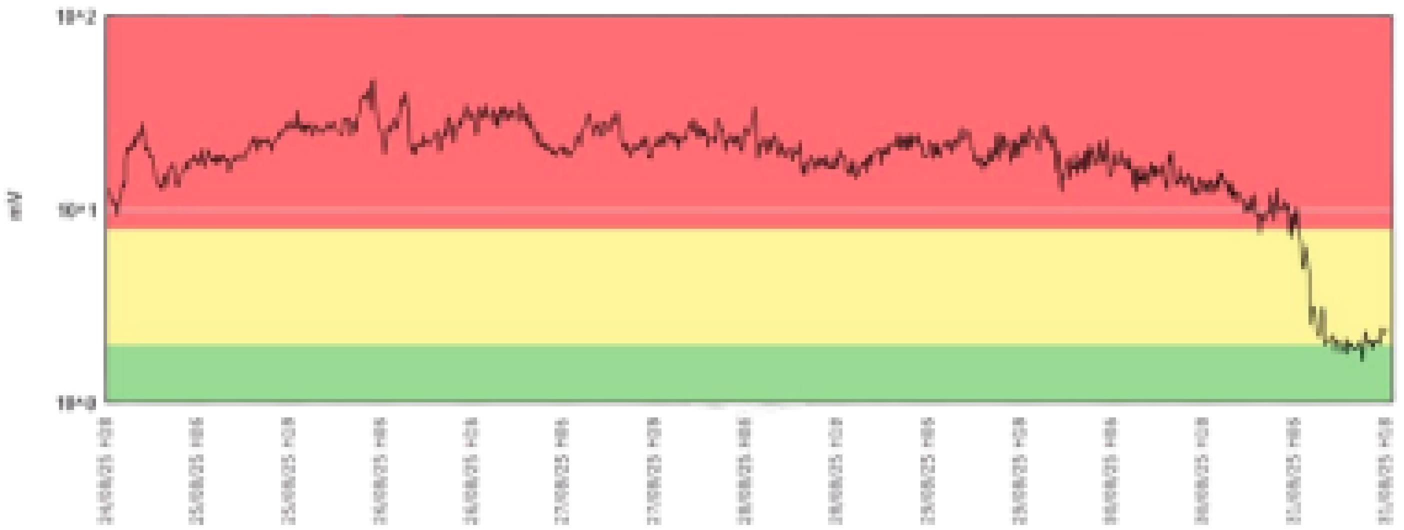
It's important to know how is structured the volcano's building; its divided in three parts:

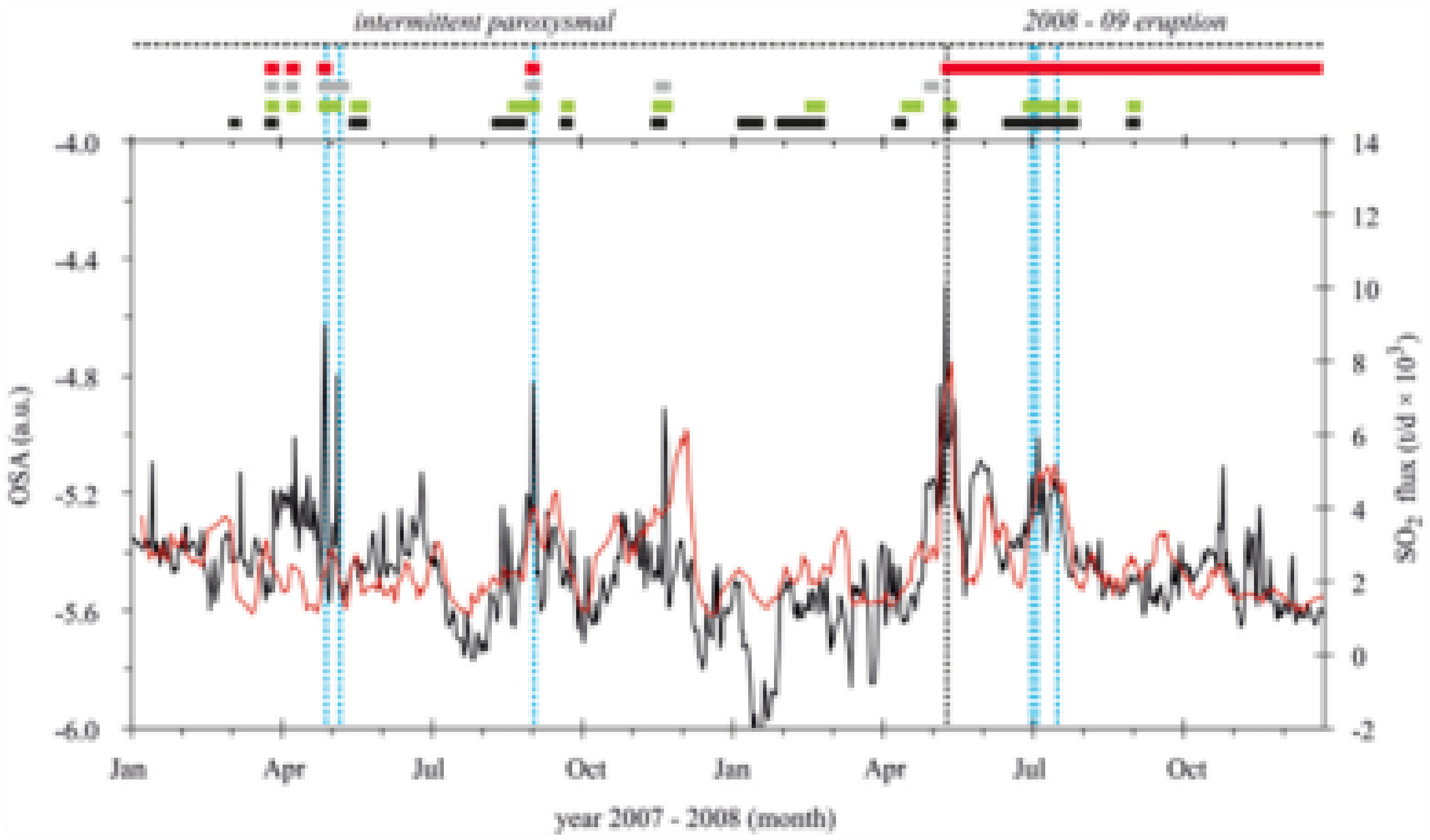
- **Deep pumbling system**: Where Magma forms under high pressure and temperature and Incorporates deep volatiles (He, C, O, S, H).
- **Intermediate pumbling system**: In which Magma rises and may accumulate in crustal reservoirs; here gases start to separate.
- **Shallow system**: In the highest part of the volcano's building it happen Final gas separation and formation of gas phases.



1. *Image on the right side is an example of a mapping visualization of the earthquake; implemented by INGV experts to localize seismic events.*

2.1 RMS Graphical representations





2. Just to have an idea, here an example of RMS graphical representation. This is how changes on volcanic tremor are typical represented at INGV.

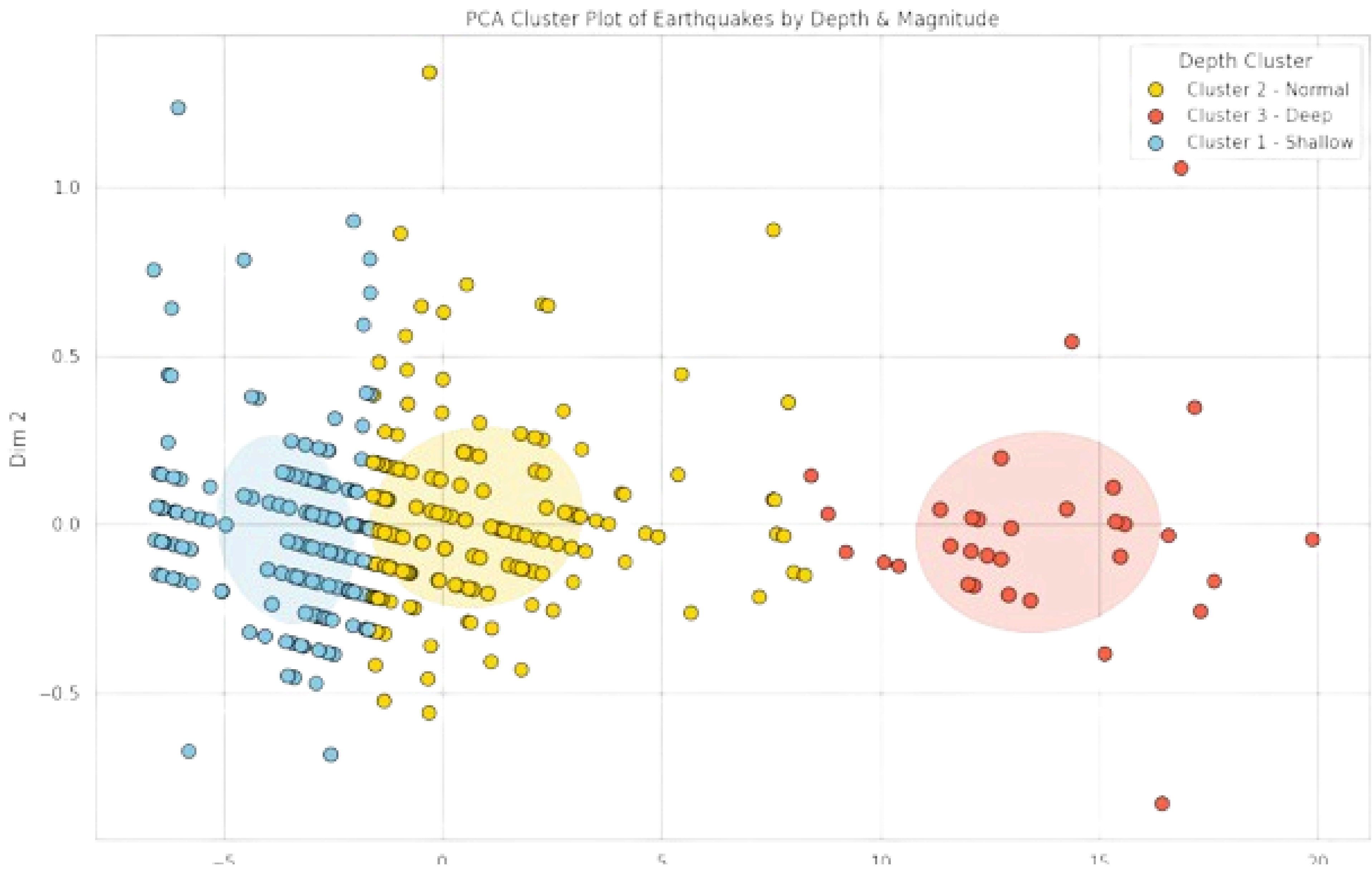
3. Data: How are collected and how to clean them

- The dataset is a CSV file with sensor data from Mount Etna.
- **Data issues**
  - Each variable required a specific cleaning approach.
- **Cleaning steps**
  - Outliers were filtered with the Interquartile Range (IQR) method and Features were standardized using Z-score and Min–Max normalization.
- **Feature engineering**
  - A 30-step sliding window was applied to add temporal context.

4.Feature selection and analysis

Some functions to see feauture's patterns:

- A **correlation matrix**:highlighting relevant variables
- **Clustering process**: Dividing earthquakes in categories based on depth (deep, intermediate, shallow)
- **PCA**: to reduce variables dimensions from three to two
- A **scatter plot:** for a better graphical visualization



3. Image represents graphical visualization after the clustering process implementation; we can distinguish earthquake in differents clusters based on their depth

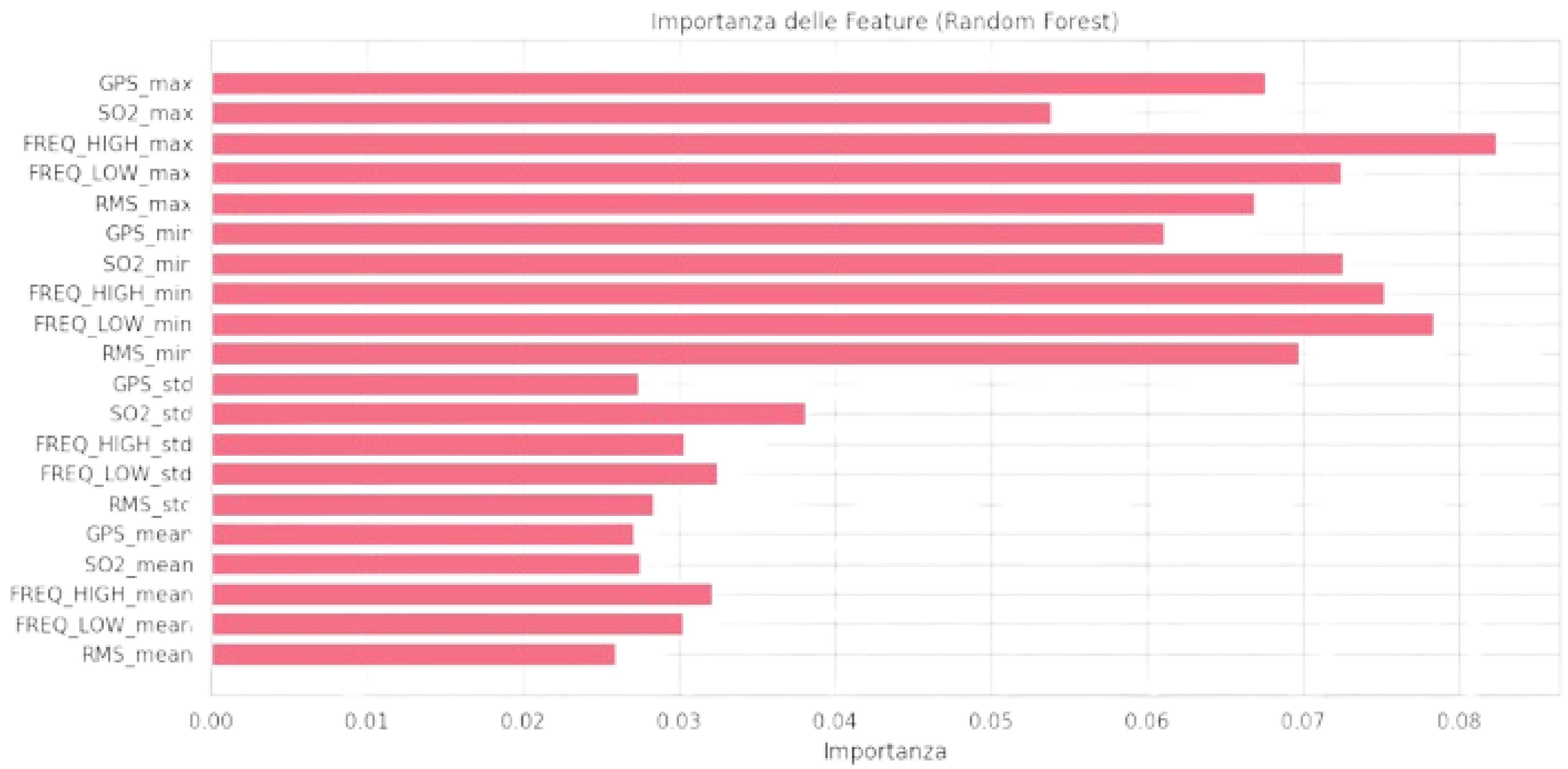
5. Train the model: Prevision of seismic events

5.1 Random Forrest

As a first attempt, I implemented a Random Forest model, which combines multiple decision trees to make predictive decisions. I carried out two different experiments:

- **Experiment 1**: Using raw data without any preprocessing; the accuracy was low, with many missed earthquake events.
- **Experiment 2**: Introducing a more robust preprocessing phase to improve performance.

5.1 Graphical visualization of the results



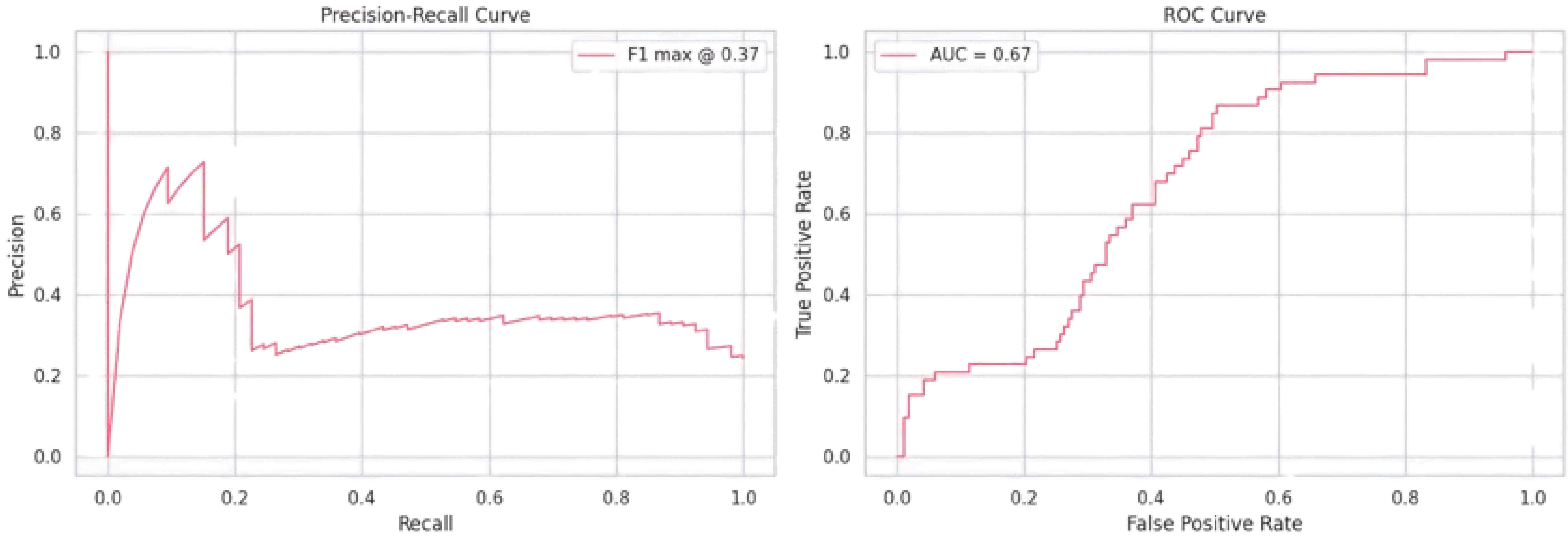
4. *The feature importance plot highlights that in this casethe most informative variables are those exhibiting sudden changes.*

5.2 LSTM Model

The analysis focused on predicting the occurrence of earthquakes within the next 24 hours by constructing a binary target variable (quake_next24h).

- The predictive model was designed with:
- a single LSTM layer (64 units)
- a dropout layer to reduce overfitting
- a dense layer (32 neurons)
- a sigmoid output layer for probability estimation.

5.2 LSTM Graphical representation



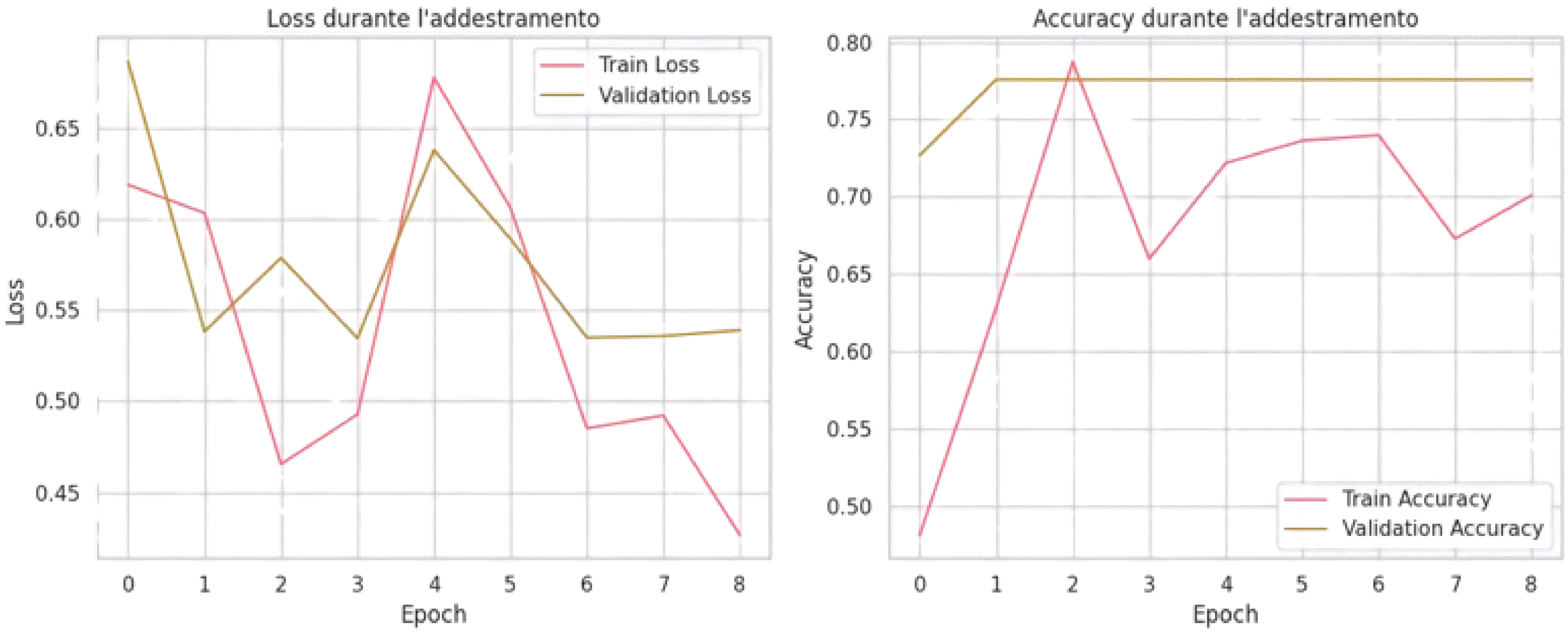
5. In the left graph we can see a high precision which higlight difficulty of the model to capture false negatives. In the ROC curve performances increases but still with imperfections; this could be a good baseline model.

5.3 Mulivariate LSTM

In natural phenomena is needed to take in to account all the variables simultaneosly;to overcome these shortcomings, a multivariate LSTM model was developed.

- **Model design**: The multivariate framework takes into account more variables (CO_2,SO_2,Clinometry)
- **Evaluation strategy**: The univariate model doesn’t captured rare events; now we focus on F1-score and recall

5.3 Multivariate LSTM graphical representations



6. *This is the proof that the model perform better taking into account rare events; but there is still an instability given by the "nature" of phenomena*

6. Reviewing the results

The performance are “theoretically” good but he model proposed could be only a simulation or baseline for deepest architeture; this is for the imprevedibility of natural phenomena and earthquake in general.

	date	ten_batt	temp_CR10	tilt_x_Avg	tilt_y_Avg	temp_tilt	nord_tilt	barometro
479203	2022-12-31 22:45:00	12.42	11.76	102.644	255.606	7.080	230.5	854
479204	2022-12-31 23:00:00	12.43	11.69	102.637	255.507	7.078	230.5	854
479205	2022-12-31 23:15:00	12.44	11.62	102.629	255.438	7.077	230.5	854
479206	2022-12-31 23:30:00	12.42	11.55	102.625	255.394	7.075	230.5	854
479207	2022-12-31 23:45:00	12.41	11.48	102.629	255.430	7.084	230.5	854

7. Here the variables taken into account during LSTM multivariate training

7. Conclusions

In conclusion, this work emphasizes the power of deep-learning architetures and models even in critical phenomena such as earthquakes; but the imprevedibility of the events make imperfect the model.

Thanks for the attention