Sapienza
Università di Roma

Department of Mathematics
Degree course in Mathematical Sciences for Artificial Intelligence

Machine Learning to Extract Predictive Information from the Pollino Seismic Catalog

Earthquake Physics & Machine Learning
Final Project

Professor:
Professor Chris Marone

Student:
Mariagiusi Nicodemo

Academic Year 2025/2026

Contents

Abstract

In this study, I investigate the predictive information content of the Pollino seismic catalog using a *machine learning* framework oriented toward *seismic activity regimes*. The study area, which has been relatively understudied in the Italian context, is characterized by complex geodynamics and strong temporal variability, making the problem intrinsically challenging. The analysis relies on homogeneous seismic data available from the late 1990s to the most recent period, while earlier sequences from the 1980s are not considered due to heterogeneity among catalogs and differences in acquisition procedures. The catalog is aggregated into weekly windows and described through interpretable features derived from recent seismicity, including multi-scale event rates and their logarithmic transforms, magnitude statistics, the fraction of events above the magnitude of completeness, and b-value estimates. High-activity weeks are defined as those exceeding the 75th percentile of weekly event counts (threshold: > 4 events/week), and a one-week-ahead prediction is performed using a time-shifted target, with a strictly time-ordered train/test split to prevent information leakage. Linear and tree-based classifiers are compared and evaluated using metrics suited to imbalanced classification. The results indicate a moderate yet statistically meaningful predictive skill, with peak ROC-AUC values around 0.83, and highlight the dominant contribution of rate-related features. Overall, the analysis suggests that the extractable predictive signal is largely rate-driven, while the inherent complexity and partial stochasticity of the seismic process constrain further performance improvements. This study explores the applicability of machine learning approaches to the analysis of seismic activity regimes in a complex tectonic setting of southern Italy.

1 Introduction

1.1 Background and Motivation

Seismic catalogs represent a fundamental source of information for investigating the temporal evolution of earthquake occurrence. However, the nonlinear, heterogeneous, and partially stochastic nature of tectonic processes makes the characterization of short-term variations in seismic activity particularly challenging. In recent years, machine learning (ML) techniques have increasingly been employed as statistical tools for the analysis of seismic catalogs, enabling the identification of temporal patterns and changes in the activity state of a system without imposing strong assumptions of stationarity.

Within this framework, a particularly suitable perspective is the analysis of *seismic activity regimes*, defined as time intervals characterized by persistently different levels of seismicity. Such regimes may include phases of elevated activity associated with prolonged sequences or swarm-like behavior, which are commonly observed in complex tectonic environments. The investigation of these patterns requires a careful treatment of temporal causality in order to prevent information leakage and preserve the physical interpretability of the results.

1.2 Objectives of the Study

The Pollino region, represents a particularly relevant setting for the analysis of seismic activity regimes. The area is characterized by a complex tectonic framework and diffuse seismicity, often organized in prolonged sequences and swarm-like episodes. In this work, I analyze the Pollino seismic catalog using homogeneous data available from the late 1990s to the most recent period. The catalog is aggregated into weekly time windows and described through interpretable features derived exclusively from past seismicity, including multi-scale seismic rates, magnitude statistics, and Gutenberg–Richter parameters. The objective is to assess whether historical seismicity can be used to probabilistically classify the activity regime of the subsequent week. The conceptual workflow of the analysis is schematically illustrated in Figure 1, highlighting its strictly sequential and time-aware structure.

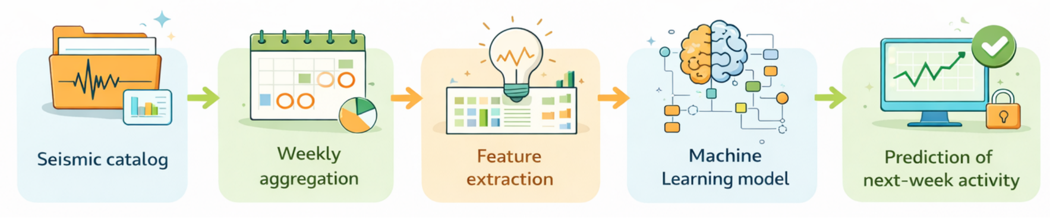


Figure 1: Schematic overview of the proposed workflow, from seismic catalog aggregation to weekly seismic activity classification.

2 Dataset and Data Handling

2.1 Seismic catalog and study area

The analysis is based on the seismic catalog of the Pollino region in southern Italy, retrieved through INGV web services. The study area is characterized by diffuse seismicity and a complex tectonic setting, with frequent swarm-like sequences. A geographic overview of the analyzed region and the spatial distribution of seismic events are shown in Figure 2.

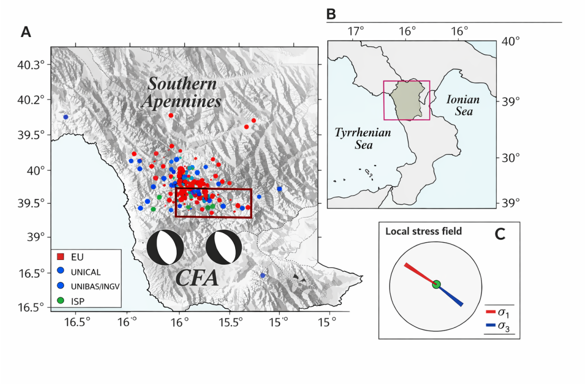


Figure 2: Geographic location of the Pollino study area in southern Italy. The outlined region indicates the spatial extent of the seismic catalog analyzed in this study.

The selected catalog includes origin time and magnitude information for each event and represents the basis for all subsequent processing steps.

2.2 Temporal aggregation and preprocessing

To formulate the problem on a regular time, the event-based catalog is aggregated on a weekly basis. Weeks with no recorded events are explicitly retained to preserve temporal continuity and allow consistent computation of rolling-window features. Figure 3 illustrates the temporal evolution of the weekly number of seismic events, highlighting both background seismicity and periods of elevated activity. Basic preprocessing steps are applied to ensure temporal causality and data integrity. Missing values arising from rolling-window operations at the beginning of the time series are handled using a time-aware strategy, and all preprocessing operations are performed without incorporating information from future observations.

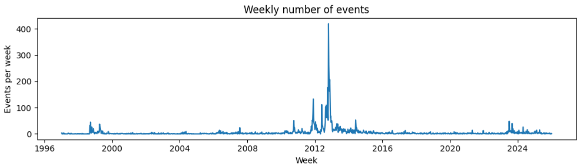


Figure 3: Weekly number of seismic events over time.

3 Feature Engineering

3.1 Rate and magnitude-based features

The most basic descriptors of seismic activity are derived from event counts and magnitude statistics computed on a weekly basis. These include the number of events per week and summary statistics of the recorded magnitudes, such as the mean and maximum magnitude. To reduce sensitivity to short-lived fluctuations and to emphasize persistent changes in activity, the seismic rate is further filtered using rolling averages computed over multiple temporal scales. In particular, rolling windows of 4 and 12 weeks are considered, allowing the model to capture variability at different characteristic time scales and potentially associated with distinct seismic regimes.

3.2 Gutenberg–Richter parameters

In addition to rate-based features, parameters derived from the Gutenberg–Richter frequency–magnitude relation are incorporated. The Gutenberg–Richter law describes the relationship between earthquake magnitude and cumulative event frequency as

$$\log_{10} N(M) = a - bM, \tag{1}$$

where $N(M)$ denotes the number of events with magnitude greater than or equal to M, and a and b are empirical constants. The b-value is of particular interest, as it provides information on the relative proportion of small to large earthquakes and is often interpreted as an indicator of stress conditions and fault heterogeneity. In this study, b-values are estimated over rolling temporal windows using a maximum-likelihood formulation,

$$b = \frac{\log_{10} e}{\overline{M} - M_c}, \tag{2}$$

where $\overline{M}$ is the mean magnitude within the window and M_c denotes the magnitude of completeness. Rolling-window estimation of the b-value allows its temporal evolution to be tracked and enables the inclusion of a physically interpretable feature that may reflect changes in the underlying seismic regime.

3.3 Feature summary and temporal consistency

All features are computed using only information available up to the current time step, ensuring strict temporal causality. Rolling-window features are aligned such that no future observations contribute to the feature values associated with a given week, preventing information leakage. The resulting feature set provides a compact and causally consistent representation of seismic activity.

4 Target Definition and Experimental Setup

4.1 Definition of seismic activity regimes

The prediction task addressed in this study is formulated as a binary classification problem aimed at identifying short-term regimes of seismic activity. The target variable represents whether the level of seismic activity in a given future time window exceeds a predefined reference threshold. Specifically, seismic activity is evaluated on a weekly basis, and the target label associated with week t corresponds to the activity level observed during week $t + 1$. For each week t, input features are computed using only information available up to that week, while the associated label indicates whether the number of seismic events in week $t + 1$ belongs to a high- or low-activity regime. High-activity regimes are identified by applying a quantile-based threshold to the weekly event counts. The threshold is computed exclusively on the training subset in order to avoid any influence from future data. Weeks with event counts exceeding this threshold are labeled as high-activity, while all remaining weeks are assigned to the low-activity class.

4.2 Train–test split and temporal causality

To preserve the intrinsic temporal structure of the seismic catalog, the dataset is divided into training and test subsets using a strictly chronological split. The earliest portion of the time series is assigned to the training set, while the most recent segment is reserved for testing. No random shuffling is applied at any stage of the data splitting process. All preprocessing steps, including feature computation, handling of missing values, and threshold selection for target definition, are performed in a time-aware manner. Statistics required for preprocessing are computed exclusively on the training data and subsequently applied to the test set. This strategy enforces strict temporal causality and prevents information leakage from the test period into the training process.

4.3 Machine learning models and evaluation metrics

Multiple machine learning models are considered to assess the predictive value of the engineered features and to evaluate the trade-off between model complexity and interpretability. Model performance is evaluated on the held-out test set using standard classification metrics, including precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). This evaluation framework allows for a consistent comparison between models under realistic forecasting conditions.

5 Results

5.1 Model performance

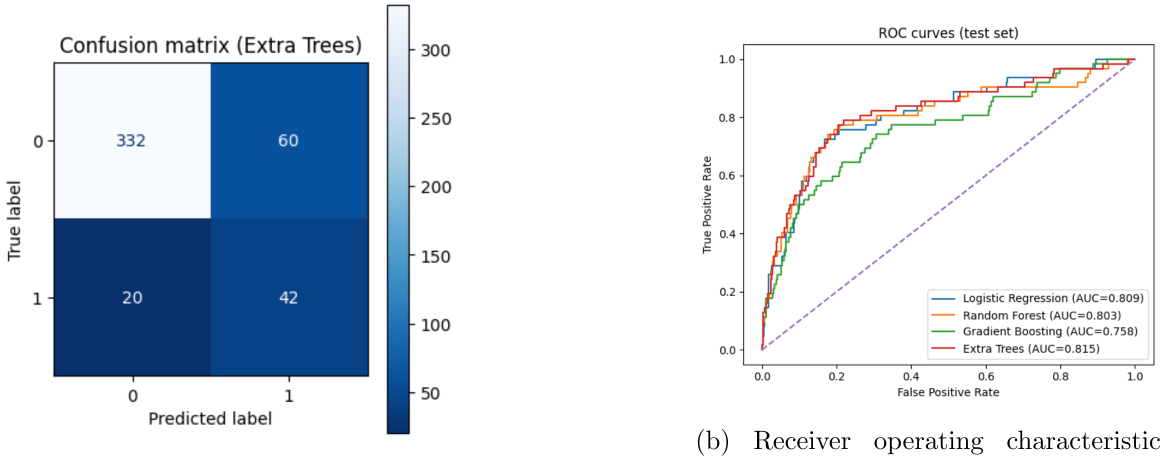
Model performance is quantified using standard classification metrics, including precision, recall, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC). These metrics provide complementary perspectives on model behavior and are particularly informative in the presence of class imbalance between high- and low-activity regimes. Table 1 summarizes the performance obtained by the evaluated models on the test set in terms of ROC-AUC. Logistic regression provides a solid and interpretable baseline, while tree-based ensemble models achieve overall better performance, suggesting the presence of nonlinear relationships between the engineered features and the target variable. Among the tested models, the Extra Trees classifier yields the highest ROC-AUC, indicating an improved ability to discriminate between high- and low-activity seismic regimes.

Table 1: Performance of the evaluated machine learning models on the test set.

Model	ROC-AUC
Logistic Regression	0.809
Random Forest	0.803
Gradient Boosting	0.758
Extra Trees	**0.815**

5.2 Confusion matrix and ROC analysis

To further analyze classification behavior, the confusion matrix and the ROC curves of the evaluated models are examined. The confusion matrix of the best-performing model, shown in Figure 4a, provides insight into the balance between correctly and incorrectly classified weeks, highlighting the model's capability to identify high-activity periods while limiting the number of false positives. The Extra Trees classifier correctly identifies a substantial fraction of high-activity weeks, achieving a recall of approximately 0.68 for the high-activity class, at the cost of a moderate number of false positives. This behavior reflects a trade-off between sensitivity to elevated seismic activity and classification conservatism, which is appropriate in a short-term regime identification context. Figure 4b shows the ROC curves for the different models evaluated on the test set. All models exhibit a clear separation from random classification, confirming the presence of exploitable predictive information in the data. Consistent with the quantitative metrics reported in Table 1, the Extra Trees model displays the highest ROC curve.



(a) Confusion matrix for the best-performing model.

(b) Receiver operating characteristic (ROC) curve for the best-performing model.

Figure 4: Classification performance of the best-performing model evaluated on the test set.

5.3 Feature importance

To investigate which features contribute most strongly to model predictions, feature importance scores are analyzed for the best-performing ensemble model. As shown in Figure 5, variables related to seismic rate dominate the ranking, particularly rolling averages and their logarithmic transformations computed over longer temporal windows. Magnitude-based features and Gutenberg–Richter parameters provide a secondary but non-negligible contribution, indicating that information related to magnitude distribution complements rate-based descriptors. Overall, the results suggest that predictive performance is primarily driven by persistent variations in seismic rate, with additional features contributing to a more complete characterization of seismic activity regimes.

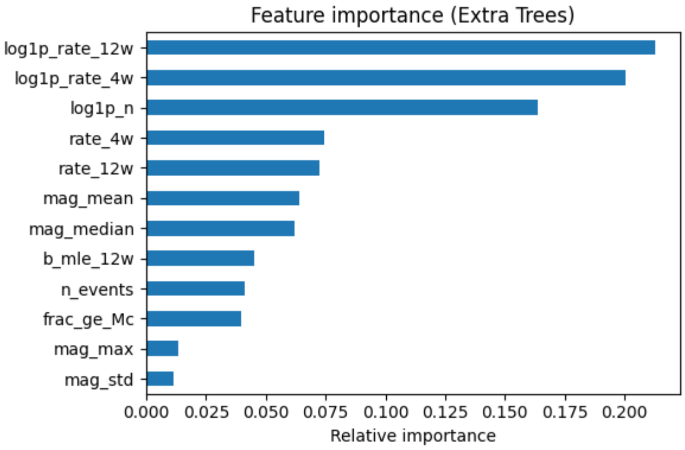


Figure 5: Feature importance for the ensemble model, highlighting the relative contribution of different seismic features.

6 Discussion and Limitations

In this study, I explored the use of machine learning techniques for the probabilistic classification of short-term seismic activity regimes based on a regional seismic catalog aggregated on a weekly basis. The results indicate that simple and physically interpretable information, particularly features related to seismic rate and its temporal evolution, contains a meaningful predictive signal for distinguishing between low- and high-activity weeks. The comparison between linear models and ensemble-based classifiers shows that the latter achieve superior performance, suggesting the presence of nonlinear relationships between the input variables and future activity regimes. However, the analysis of feature importance clearly highlights that the dominant contribution to the predictive signal is provided by rate-related variables, especially rolling averages computed over longer temporal windows and their logarithmic transformations. In order to assess the robustness of the obtained results, I conducted additional analyses using more advanced modeling configurations. These included preliminary hyperparameter exploration for ensemble classifiers and alternative time-aware validation strategies designed to further stress-test the predictive framework. Importantly, the overall behavior of the models and the relative contribution of the input features remained consistent across all tested settings, with rate-related variables continuing to dominate the predictive signal. An additional robustness analysis was performed by varying the percentile threshold used to define high-activity seismic regimes. In particular, thresholds corresponding to the 70%, 75%, and 80% quantiles of the weekly event count distribution were considered. The results show that model performance, as measured by the ROC-AUC, remains stable and on the order of 0.8 across all tested thresholds, indicating that the predictive signal is not critically sensitive to the specific definition of the target variable. As the threshold increases, an expected trade-off between precision and recall for the minority class is observed, consistent with the increasing class imbalance. Overall, these findings confirm the robustness of the main conclusions with respect to moderate variations in the definition of activity regimes. It is important to emphasize that the obtained performance should be interpreted in light of the intrinsic limitations of the problem. Seismicity is a complex and partially stochastic process, and the amount of predictive information that can be extracted from historical catalogs is inherently limited. In this context, the lack of substantial performance gains achieved by increasing model complexity suggests that the primary constraint is not algorithmic, but rather related to the information content of the seismic catalog itself. Moreover, the present analysis relies exclusively on temporally aggregated information and does not incorporate spatial features or additional physical proxies, which may provide further insights in future studies.

7 Conclusions and Future Work

In this work, a machine learning framework for the probabilistic classification of short-term seismic activity regimes was developed and evaluated using a time-aware experimental setup. By aggregating seismic catalogs on a weekly basis and extracting physically motivated features, the proposed approach demonstrates that meaningful information on future activity levels can be inferred from past seismicity.

The results show that ensemble machine learning models outperform a simple baseline classifier and are able to capture temporal patterns associated with elevated seismic activity. Importantly, the study emphasizes methodological rigor, particularly with respect to temporal causality and information leakage, ensuring that the reported performance reflects realistic forecasting conditions.

Future work may extend this framework in several directions. These include the application of the methodology to multiple seismic regions, the exploration of alternative temporal aggregation scales, and the incorporation of additional features derived from spatial information or stress-related proxies. Furthermore, integrating physics-informed constraints or hybrid modeling approaches may improve both interpretability and robustness.

Overall, this study supports the role of machine learning as an exploratory and complementary tool in earthquake physics, capable of aiding the characterization of seismicity patterns and contributing to a probabilistic understanding of short-term seismic activity.

Appendix

This appendix reports a supplementary analysis aimed at assessing the sensitivity of the results to the choice of the threshold used to define high-activity seismic regimes. As discussed in Section 6, this analysis is not intended to introduce new primary results, but rather to document the robustness of the adopted predictive framework.

Specifically, I considered three different percentile thresholds of the weekly event count distribution, corresponding to the 70%, 75%, and 80% quantiles. For each threshold, the Extra Trees classifier was retrained and evaluated on the same test set, while keeping all other methodological choices unchanged. Table 2 reports the main evaluation metrics for the minority class (high activity), including ROC-AUC, precision, recall, and F1-score.

The results indicate that overall model performance remains stable across the tested thresholds, with ROC-AUC values on the order of 0.8 in all cases. The observed variations in precision and recall reflect the expected trade-off associated with different levels of class imbalance. These findings support the conclusions discussed in the main text and confirm that the choice of the 75% percentile as the reference threshold does not critically affect the overall interpretation of the results.

Table 2: Robustness analysis of the target definition using different percentile thresholds. Performance metrics are reported for the minority class (high-activity regime).

Percentile	**Threshold**	**Class Balance**	**ROC-AUC**	**Precision**	**Recall**	**F1-score**
70th	3.0 events/week	23% (98/438)	0.760	0.504	0.684	0.580
75th	4.0 events/week	14% (62/438)	0.820	0.409	0.726	0.523
80th	5.0 events/week	10% (46/438)	0.807	0.340	0.717	0.462

References

[1] Beno Gutenberg and Charles F. Richter. Frequency of earthquakes in california. *Bulletin of the Seismological Society of America*, 34:185–188, 1944.

[2] Keiiti Aki. Maximum likelihood estimate of b in the formula log n = a - bm and its confidence limits. *Bulletin of the Earthquake Research Institute*, 43:237–239, 1965.

[3] IRIS DMC. Fdsn web services. *Seismological Research Letters*, 86(6):186–190, 2015.

[4] Thibaut Perol, Moussa Gharbi, and Marine Denolle. Convolutional neural network for earthquake detection and location. *Science Advances*, 4(2):e1700578, 2018.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[6] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An introduction to statistical learning. *Springer Texts in Statistics*, 2021.

[7] Chris Marone. Laboratory-derived friction laws and their application to seismic faulting. *Annual Review of Earth and Planetary Sciences*, 26:643–696, 1998.