**ENGLISH**

**Topic/Title**

***Development of Machine Learning – based models to optimize the analysis of multidimensional geological datasets for the reconstruction of complex geological processes***

**Proposer (Tutor)**

***Prof. Ester PIEGARI***

**Research proposal**

The main scientific question underpinning this PhD proposal is whether and how machine learning techniques can optimize the information extracted from scientific deep ocean drilling and to contribute to the reconstruction of complex geological phenomena such as the evolution of deep water circulation and its interaction with sedimentary processes along the continental margins.

The undeniable increase in the volume and complexity of geological data available nowadays highlights the importance of machine learning (ML) techniques, not only to extract as much useful information as possible but also to gain new insights from the data (Bergen et al., 2019). Some ML algorithms can be used to automate the analysis of large and complex datasets facilitating classification tasks, and enabling a deeper understanding of the natural Earth processes by uncovering new relationships and hidden patterns within the data (Bhattacharya & Di, 2022; Uddin et al., 2022). In particular, unsupervised learning is especially suitable for exploratory data analysis and visualization in multidimensional spaces. Several clustering approaches have demonstrated the potential to group diverse types of data, identifying patterns and structures without prior labeling, thus facilitating the identification of geological features (La Marca & Bedle, 2022;

Piegari et al., 2023; Kumar et al., 2024). Dimensionality reduction techniques applied to clusters in high-dimensional spaces simplify complex datasets while retaining essential information, simultaneously suggesting hierarchies among parameters to focus on. Since unsupervised learning approaches can reveal unexplored connections within data, their use is crucial and promising for identifying teleconnections, as these patterns often manifest as correlations across different geographical locations and time periods.

The aim of the project is precisely to explore unsupervised approaches to improve the analysis of large multidimensional datasets, particularly related to scientific ocean drilling data, to optimize the extraction of useful information for reconstructing geological processes (Rodrigues et al., 2022). Specifically, the main objectives of the project can be summarized as follows:

- Data Integration: to collect and to integrate diverse geological datasets derived from scientific ocean drilling activities, including lithological, geophysical, and geochemical information.
- Model Development: to develop machine learning models aimed at classifying geological features and reconstructing the geological history through the identification of patterns and correlations in the data.
- Performance Evaluation: to analyze the effectiveness of various machine learning algorithms in terms of accuracy and computational efficiency.

**References**

Bergen KJ et al (2019) Machine learning for data-driven discovery in solid Earth geoscience. Science, 363, 6433

Bhattacharya S & Di H (2022) Advances in Subsurface Data Analytics, Elsevier (2022)

Kumar PC et al (2024) Unsupervised learning approach for revealing subsurface tectono-depositional environment: A study from NE India. Journal of Applied

Geophysics, 229, 105478

La Marca, K & Bedle H (2022) Deepwater seismic facies and architectural element interpretation aided with unsupervised machine learning techniques: Taranaki Basin, New Zealand. Marine and Petroleum Geology, 136, 105427

Piegari E et al (2023) A machine learning-based approach for mapping leachate contamination using geoelectrical methods. Waste Management, 157, pp. 121 – 129

Rodrigues, S, Hernández-Molina, FJ, Larter, RD, Rebesco, M, Hillenbrand, CD, Lucchi, RG, Rodríguez-Tovar, FJ (2022) Sedimentary model for mixed depositional systems along the Pacific margin of the Antarctic Peninsula: Decoding the interplay of deep-water processes,
Marine Geology, 445, 106754

Uddin S., et al. (2022) Machine learning in project analytics: a data-driven framework and case study. Sci Rep 12, 15252

**Research Plan**

***I° year***

**Data Collection and integration:** state-of-the-art review; acquisition and organization of multidimensional geological datasets derived from scientific drilling, including lithological, geophysical, and geochemical data; preprocessing to ensure data homogeneity and quality.

**Training and initial development of ML models:** study and implementation of machine learning algorithms, with particular focus on unsupervised learning; preliminary experimentation on test datasets to validate methodologies.

**Conference participation and networking:** to attend and possibly to present a poster and/or talk to at least one national or international conference; begin

planning the research period abroad.

***II° year***

**ML model development and data analysis:** development of machine learning models to classify geological features and identify patterns and correlations in multidimensional data.

**Research period abroad:** stage at a research institute or foreign university specializing in machine learning applied to geosciences, for collaboration and advanced training.

**First scientific publication:** preparation and submission of the first peer-reviewed article focused on methodologies and initial results obtained with ML models applied to the datasets used.

***III° year***

**Performance Evaluation and Optimization**: comparative analysis of the effectiveness of the algorithms used; optimization of models based on results and feedback received.

**Second scientific publication and dissemination**: preparation and submission of a second scientific article; presentation of the results at international conferences.

**Thesis writing**: finalization of the PhD thesis integrating all research phases and related results.