**Cerinte Laborator 1**

**(10 puncte)**

* Membri:
  + Mara Vălean
  + Tania Săsăran
  + Jaclina-Iana Bulat
  + Bianca Szekely
  + Mark Doszlop
* Stabilire denumire echipa:
  + QORn
* Identificare si discutie roluri membri echipa (roluri related to AI):
  + Data Engineers: Jaclina, Mark (data preprocessing, data preparation)
  + Data Scientists: Tania, Mara, Bianca (AI algorithm, AI tools)
* Identificare tema pe care va lucra echipa:
  + Identificare corelații între Radon, CO2 și cutremure
* Identificarea a 2 articole relevante din literatura de specialitate pe tema aleasa de catre echipa:
  + a)
  + b) Analysis of 7-years Radon time series at Campi Flegrei area (Naples, Italy) using artificial neural network method F. Ambrosino a b, C. Sabbarese a b, V. Roca b, F. Giudicepietro c, G. Chiodini
* Articol: [**Analysis of 7-years Radon time series at Campi Flegrei area (Naples, Italy) using artificial neural network method**](https://www.sciencedirect.com/science/article/abs/pii/S0969804320303894?casa_token=1rOeacPGMPEAAAAA:J5-piup9G6TLrSvA2PUCFw0OwFPCNZ1YM_a9eLbQx3WUIEQUc52XTYoAcI5FcMpRT9QQhvIm-S4)
  + - setul de date folosit: set de date creat de ei
      * descriere: data recorded over 7 years in the volcanic caldera of Campi Flegrei (Naples-Italy) from 1 july 2011 until 31 december 2017
      * disponibilitate: indisponibil
      * tipul datelor: set de date numeric tabelar
    - algoritmii folositi: Artificial neural network (ANN) method
      * descrierea algoritmilor inteligenti folositi: An ANN is a computational method inspired by the operation of the biological neural system, which predicts the unknown values by processing the basic information characterizing a system, such as the biological nerve cell. In this study, the Radon activity concentration time series is estimated by ANN method.
      * parametrii: The input layer is composed by 6 neurons (parameters): the fumarolic tremor, the background seismicity, the CO2 gas concentration, the temperature, the relative humidity and the pressure.
    - metricile calculate
      * identificare: Pearson’s coefficient – ρ,
      * intelegere: the linear correlation between the measured and the trained signals
    - rezultatele obtinute:
      * The result shows that the distribution of Radon activity concentration obtained by neural network is well estimated; the correlation coefficient between measured and replicated signal is 0.82

Article b)

Anomaly Classification For Earthquake Prediction In Radon Time Series Data Using Stacking And Automatic Anomaly Indication Function

* Adil Aslam Mir, Fatih Vehbi C¸ Elebi, Muhammad Rafique, M. R. I. Faruque, Mayeen Uddin Khandaker, Kimberlee Jane Kearfott, And Pervaiz Ahmad

1. Dataset

Description:

* Objective: The dataset is used to categorize the soil radon gas concentration into seismically active and non-active time series. The main goal is to predict seismic activities by analyzing anomalies in SRG concentrations.
* Location: The SRG data was collected from a fault line in Muzaffarabad, a city in the Pakistan territory of Kashmir.
* Time Period: The data collection spanned from March 1, 2017, to February 28, 2018.
* Seismic Events: During the study period, four seismic events were recorded, which are considered in the analysis for earthquake prediction.

Data Types and Attributes:

* Radon Concentration: The primary focus of the dataset is on the concentration of soil radon gas, measured in becquerels per cubic meter (Bq/m³).
* Environmental Parameters: Temperature, pressure, and humidity are also considered to analyze their influence on radon emission and to better understand the anomalies related to seismic activities.
* Seismic Activity Labels: The dataset includes manual labeling of radon concentration data as seismic and non-seismic based on seismic events recorded by the Pakistan Meteorological Department.

Data Collection Methodology:

* Instrumentation: An RTM 1688-2 (SARAD RTM 1688-2, Nuclear Instruments, Germany) instrument was used to measure the SRG concentration.
* Frequency of Measurements: Each reading was taken at intervals of 40 minutes, resulting in 36 overall readings for one complete day.

Availability:

* The dataset comprises soil radon gas concentration measurements collected by the RTM 1688-2 instrument for specific research on earthquake prediction. However, the article does not clarify whether this collected data is accessible to the public. Since the data was gathered using specialized equipment and for scientific analysis, its availability for broader use or external research purposes is not specified within the document.

1. Used algorithms

Description: The used algorithms involve a two-layer methodology combining a stacking ensemble approach with an Automatic Indication Function (AAIF) for post-processing.

1. First Layer: Stacking Ensemble-Based Approach

* Description:

This approach integrates predictions from three base learners to train on seismically active and inactive periods for predicting Soil Radon Gas (SRG) concentration. These predictions are then combined with the labeled anomaly data used to train a meta-learner that categorizes the data into active and non-active radon time series.

* Base Learners:

- *Generalized Linear Model (GLM) - Learner1*: Used for predicting SRG concentration based on input features related to Non-Seismically Active Data (NSAD).

- *Linear Regression- Learner2:* Predicts SRG concentration, focusing on the relationship between seismic activities and SRG concentration.

- *K-Nearest Neighbors (k-NN)- Learner3*: Utilizes seismically active data to classify SRG data into seismically active or inactive categories.

* Meta-Learner:

- *Support Vector Machine (SVM) with a Radial Kernel*: This meta-learner is trained with the combined predictions of the Base Learners. It categorizes radon concentration into

seismic and non-seismic concentrations based on the labeled anomaly data. These predictions serve as the final classifications for the unseen data.

* Parameters: During the training phase, radon concentration in soil gas was learned from measured environmental data (i.e., relative humidity, temperature, and pressure) using three learners. From these three learners, radon concentration of validation data is predicted, and a Meta feature matrix is calculated, where predicted radon concentration acts as Meta features along with class labels. The Meta feature matrix is learned using another machine learning method (support vector machine with a radial kernel) with the initial labelling of radon gas time series data relying on that Meta learner.

1. Second Layer: Automatic Anomaly Indication Function (AAIF)

* Description:

In the second layer, the classifications produced by the stacking ensemble are passed to an AAIF. This function labels the time series into seismically active (1) or non-active (0) based on Soil Radon Gas (SRG) concentration predictions. Initially, class labels are assigned to data, which the AAIF then evaluates. If the percentage of seismic events identified exceeds a set threshold, the day's data is marked as indicative of seismic activity.

* Parameters: The indication factor, set to 0.55, determines the threshold above which a full day's samples are considered anomalous and assigned a class label of 1 (seismically active).

A diagram of a process

Description automatically generated

A diagram of a model

Description automatically generated

1. Metrics

The article encapsulates several known metrics to evaluate the performance of the algorithm.

1. Accuracy:

* Mentioned for different window sizes (0-3) in the stacking ensemble approach.
* Training accuracy was considered, with values of 0.998, 0.977 for window 0 and windows 1-3, respectively.
* Testing accuracy reported values of 0.971, 0.985, 0.967, and 0.968 for windows 0-3, respectively.

1. Precision, Recall and F1

* While precision is concerned with the accuracy of positive predictions, recall focuses on capturing all actual positive instances.
* F1 Score balances precision and recall, providing a comprehensive measure of a model's performance.
* These metrics provide insights into how well the model is performing in terms of its ability to correctly identify positive and negative instances.

1. Receiver Operating Characteristics (ROC) Curve:

* Used as an effective graphical representation to analyze classification performance.
* Sensitivity (true positive rate) plotted against 1-Specificity (false positive rate).
* Area under the curve (AUC) is considered a widely used global tool for evaluating the accuracy of the test.
* AUC values provided for different window sizes (0–3) without applying the AAIF.

1. Mathews Correlation Coefficient (MCC):

* Used to assess the performance towards positive class (seismic activity).
* Noted that MCC values, especially at window 0, show that the performance towards positive class is low.

1. Anomaly Identification by Anomaly Indication Function (AAIF):

* AAIF refines the predicted class labels for a complete day into either 0 (Non-Seismic Activity) or 1 (Seismic Activity), eliminating anomalies.
* Performance metrics (testing accuracy, AUC, and precision) are presented after passing the data through the AAIF.

1. Indication Factor and Indication Percentage:

* Indication percentage is used to determine whether clumps of whole-day readings are classified as seismic.
* Usage of the indication factor (IF) and indication percentage in the AAIF.

1. Results

Results demonstrate the effectiveness of the proposed methodology in efficiently classifying radon concentration as anomalous or non-anomalous, with testing accuracies ranging from 0.985 to 0.967. The automated anomaly indication function enhances label smoothing, further smoothing the labelling.