1. The US detectors are extracted from the full dataset of safecast.org using the PIG script contained in the "Selecting and cleaning US_data" folder
   Input: measurements. csv
   Output: US_cleaned.csv
2. The file obtained is then used as input for the hadoop streaming (Hadoop version 2.6) scripts contained in the "Clustering Sensors with Hadoop" folder
   Input : US_cleaned.csv
   Output: US_results.csv
3. Using the script “getWeather.py” in the "Coupling with Weather" folder the dataset is coupled with the weather information
   Input: US_cleaned.csv
   Output: Stationary_data_with_weather.csv
4. Once the dataset is created the next step is predicting the future values of background radiation
   Using the “Model” class a single dataset is cleaned, preprocessed and the prediction could be done using different regression technique.
5. Since the inputs could be reduced with different dimensionality reduction techniques and different percentage for the training set could be tried, the script “GetBestInputs” discover which ones are the best
   Input: Stationary_data_with_weather.csv
   Outputs:
   1. SNRlinear.csv
   2. SNRrbf.csv
   3. SNRsigmoid.csv
   4. KN.csv
6. Once the best inputs are defined (best percentage of the training set and best dimensionality reduction technique) the Gridsearch technique is applied to find the best meta parameters of the support vector regression (C,gamma and epsilon) using the algorithm “GetBestParameters.py” in the corresponding folder.
   Input: Stationary_data_with_weather.csv
   Output: BestParameters

Andrea Mattera