Data Science and AI for Industrial Systems

Exam Assignment

1. Objectives

The main objective of the present assignment is to apply the knowledge learned during the course. Is requested to perform a load profile characterization analysis and an internal energy benchmarking process.

The assignment can be divided in the following phases:

- **Data preparation and visualization:** the dataset must be prepared according to the following steps:
  - Merge the different data sources (i.e. building consumption and weather) for both training and testing.
  - *Identification of statistical outliers*: statistical outliers should be identified (if they exists) according to one of the methods analyzed during the course and removed for both the training and testing dataset.
  - *Missing values replacement*: missing values should be replaced according to one of the methods analyzed during the course for both training and testing set. Pay attention to not abuse of replacement methods. Avoid filling gaps larger than 3 timesteps.
  - *Data visualization*: describe the time series data through different visualization techniques. Take inspiration from the different examples introduced during the course.
- **Load profiles characterization**: perform on the training dataset a load profile characterization process based on unsupervised clustering algorithms according to the following steps:
  - *Dataset manipulation*: the training dataset should be organized into a MxN matrix to perform clustering according to one of the methods introduced during the course.
  - *Identification of the "best" clustering solution*: the "best" number of clusters should be identified according to both evaluation metrics (i.e. silhouette index and davies bouldin index) introduced during the course and personal considerations with respect to the results obtained.
  - *Data visualization and comment*: visualizations describing the solution obtained should be produced with a critical comment of the results.
- **Energy benchmarking model:** Develop a regression model to perform a benchmarking process of energy consumption according to the following steps:
  - *Input data selection*: according to the results of the load profile characterization phase the training set CAN be filtered removing records relative to load profiles representing infrequent or anomalous consumption patterns.

- *Model training and selection*: a regression model based on one of the methods available in the scikit-library should be trained. Different methods of the scikit library can be tested (https://scikit-learn.org/stable/supervised_learning.html). If reputed necessary, hyperparameter tuning should be performed employing cross validation as introduced during the course.
- *Performance evaluation*: the performance of the model should be evaluated using appropriate metrics, such as mean absolute error, and root mean square error.
- *Model deployment*: the trained (and validated) model should be, applied to the testing dataset to estimate the energy consumption. The estimated energy consumption should be compared with the actual energy consumption to evaluate overconsumption and underconsumption as introduced during the course.
- *Data visualization and comment*: visualizations describing the solution obtained should be produced with a critical comment of the results.

- **Presentation / Deliverable:** A report (i.e. powerpoint presentation) should be prepared summarizing the steps taken to prepare the data, visualize the load profiles, and train the regression model. The report should include a description of the dataset used, preprocessing techniques applied, visualizations of the load profile data and clusters identified, analysis of the load patterns, description of the regression model used, and the comparison of estimated vs actual energy consumption during the testing period. Students are required to give a presentation of their work to the class during the final exam session.

2. Dataset description

The datasets are available in the following github folder: https://github.com/baeda-polito/aiis-energy-mllabs, in the subfolder Assignment->data.

Two typologies of dataset are present in form of .csv files:

- Building consumption timeseries: the name is defined according to the following code **building_N_weather_W**, where N is the ID of the building (from 1 to 10) and W is the ID of the relative weather file (from 1 to 4) that can be associated to the building power consumption timeseries. Power consumption data are provided with hourly frequency and are expressed in kW. For each building both training and testing datasets are provided.
- Weather timeseries data: the name is defined according to the following code **weather_W,** where W is the ID of the weather file (from 1 to 4).

3. Group assignment

The following table summarizes datasets assigned to each group of students.

Group	Datasets
Group 1	building_1_weather_1 / building_6_weather_2
Group 2	building_2_weather_1 / building_7_weather_2
Group 3	building_3_weather_1 / building_8_weather_3
Group 4	building_4_weather_1 / building_9_weather_4
Group 5	building_5_weather_1 / building_10_weather_4