iscte

INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Introduction to Machine Learning — 2025/2026

Final Project

> It is suggested that the project is done using Python notebooks due to the ability to generate a report integrated with the code. It is assumed you are proficient with programming. All answers must be justified and the results discussed and compared to the appropriate baselines. In addition to the technical report integrated with the code, intermediate datasets must also be handed-in.
> The maximum score of the project is 8 points. In the case of groups with more than one member, the report should indicate each members contribution to the work. *It is mandatory to make an oral presentation (and discussion) of the project as well as a report.*
>
> **Deadlines:**
>
> Presentation / Discussion: December 9th or 10th, 2025 (if possible, during class-time)
>
> Report: January, 15th, 2026 – 1st season (submitted via Moodle)

The objective of this project is to apply the CRISP-DM methodology to data graciously made available by the Municipality of Maia (related to energy consumption) to:

1. Characterize different consumers;
2. Test the predictive capabilities for energy time-series.

The result must have a clear graphical comparison of the modeling techniques used, displaying a clear evaluation of the potential of this data for clustering and prediction.

The energy dataset, graciously made available by the Maia Municipality, will be your main data source (soon to be available on Moodle[1]). Other available datasets (namely, those made available by E-Redes[2]) can be used as additional data. The Maia Municipality dashboard [3] can be used as inspiration for some of the visualizations.

[1] moodle.iscte-iul.pt

[2] https://www.e-redes.pt/pt-pt

[3] https://baze.cm-maia.pt/

The project and the report should follow the phases of the CRISP-DM methodology (except for Deployment, obviously).

Dataset

The dataset [D4Maia], contains $\tilde{6}$ x 10E6 records with 7 attributes. The measures are of different CMMaia buildings, that harbor municipal services. Each line is an energy reading, of a certain place at a certain time. Values are measured every 15 minutes. Each sample has the following features:

id Sample id

CPE Location code

hora Sample time

DadosDeConsumo Consumption data used for billing (kW/h)

PotAtiva Power used to calculate the consumption (kW/h)

PotReactIndut Inductive Reactive Power (VAR)

PotReactCapac Capacitive Reactive Power (VAR)

Reactive power is the result of reactive loads such as inductors and capacitors. Even though they dissipate zero power, the fact that they drop voltage and draw current gives the deceptive impression that they actually do dissipate power. This “phantom power” is called reactive power, and it is measured in a unit called Volt-Amps-Reactive (VAR), rather than watts. Reactive power may be a sign of particular types of devices in the network, but, for the time-series modeling part of this assignment, these variables (PotReactIndut and PotReactCapac) can be ignored. However, they should be present in the Data Understanding phase.

In the Data Understanding phase of CRISP-DM, take time to explore the dataset carefully. In your presentation and report, include the distributions of each variable and highlight the main correlations between them. All features except IDs and CPE should be considered in the statistical analysis. It is also important to visually explore the data, for example by showing plots that illustrate relationships between pairs of features.

For the purpose of constructing the sequence prediction models this dataset should be transformed in sequences of observations per CPE, where Consumption Data (DadosDeConsumo) and / or Active Power (PotAtiva) are the main values of interest.

For the clustering and non-sequential prediction-models, it is recommended that features are created for each CPE, that enable and facilitate the characterization of each consumer.

Experiments

You should perform the following experiments:

1. Use unsupervised learning to define and characterize sub-groups of the data, notice if any of these clusters has particular characteristics and detect possible outliers. If possible, characterize different consumers (e.g. day-services, night-services, ... etc), based on the their consumption profile. Options must be adequately justified. Suggested algorithms: K-Means and DBScan;

2. Use supervised learning algorithms to attempt to predict values for the following week, using the later 30% of each energy series as test-set, and the initial 70% as training-set. Do this in two ways, using as input:

   (a) the time-series values, per customer (suggested algorithms: ARIMA and LSTM),

   (b) feature-sets per customer (suggested algorithms: XGBoost, Random Forest and MLP).

   Notice that, you may use the results and features of clustering to guide model construction.

3. Experiment with data normalization, and analyze the differences in the results.

The feature-sets are extracted from the raw data. Notice that some of these feature-sets must encode time-related information, and can be, for example:

- avg_afternoon_peak_value,
- avg_daily_peak_time,
- avg_time_below_50%_consumption,
- ...

This is the main part of the assignment. A good set of extracted features can make quite a difference in the final result.

The technical evaluation should include different metrics to better understand the errors of the supervised machine learning approaches, as well as the clusters' quality.

The unsupervised experiments should focus on understanding groups of CPE that behave similarly, and characterizing the differences between cluster. If possible some of these profiles should be related to known profiles (detailed above).

The target for the supervised experiments should be to predict the value of attribute *Data Consumption* or *Active Power* at a certain date and time. The values used as input must be all calculated based on data *that is at least a week older* than the prediction.

The results of the prediction should be compared with a baseline that is calculated simply by assuming that the consumption at a certain time on a certain date will be the same as the consumptions the same time, a week before.

All available python libraries can be used to solve the project. The use of programming aids is, as usual, advised, although all code must be explainable by the authors during the checkpoints (during the last weeks of classes).

Evaluation

Presentation must be scheduled (scheduler will be available on Moodle). Report should be delivered on PDF and notebook (samples of the intermediate files necessary to run the notebook should be included, provided the size limit for handing in assignments on Moodle is respected). An appropriate Assignment Delivery Module will be available.

References

D4Maia. Baze. URL `https://baze.cm-maia.pt/`.