The context

Climate change is an urgent, relevant and multi-dimensional global issue with a significant impact on energy policy and infrastructure. Tackling climate change involves both mitigation (i.e. reducing greenhouse gas emissions) and adaptation (i.e. preparing for the inevitable consequences). Mitigating greenhouse gas emissions requires changes to electricity systems, transport, buildings, industry and land use.

According to a report published by the International Energy Agency (IEA), the life cycle of buildings, from construction to demolition, is responsible for 37% of global CO2 emissions linked to energy and processes in 2020. Yet it is possible to significantly reduce the energy consumption of buildings by combining easy-to-implement solutions with cutting-edge strategies. For example, renovated buildings can reduce heating and cooling energy requirements by 50-90%. Many of these energy efficiency measures also deliver overall savings and other benefits, such as cleaner air for occupants. This potential can be achieved while maintaining the services provided by the buildings.

The dataset and challenge

It is clear that accurate predictions of energy consumption for a given building as a function of its characteristics can help policy makers target renovation efforts in order to maximise emissions reductions.

The data we will be using comes from the Lawrence Berkeley National Laboratory (Berkeley Lab) and the work to be carried out consists of analysing differences in the energy efficiency of buildings in order to build one or more models to predict the energy consumption of buildings.

To do this, the data provided describes the characteristics of the buildings and the climatic and meteorological variables for the regions in which the buildings are located.

Dataset description

The dataset contains approximately 100k observations collected over 7 years in different locations on building energy use.

The dataset includes building characteristics (e.g. floor area, type of installation, etc.), meteorological data for the building location (e.g. mean annual temperature, total annual precipitation, etc.) as well as energy consumption for the building and the given year,

Each line corresponds to a single building observed in a given year.

Your task is to predict the site EUI for each row, given the building characteristics and the meteorological data for the building location.

Evaluation metrics: root mean square error

Features

- `id`: building id
- `Year_Factor`: anonymized year in which the weather and energy usage factors were observed
- `State_Factor`: anonymized state in which the building is located
- `building_class`: building classification
- `facility_type`: building usage type
- `floor_area`: floor area (in square feet) of the building
- `year_built`: year in which the building was constructed
- `energy_star_rating`: the energy star rating of the building
- `ELEVATION`: elevation of the building location
- `january_min_temp`: minimum temperature in January (in Fahrenheit) at the location of the building
- `january_avg_temp`: average temperature in January (in Fahrenheit) at the location of the building
- `january_max_temp`: maximum temperature in January (in Fahrenheit) at the location of the building
- `cooling_degree_days`: cooling degree day for a given day is the number of degrees where the daily average temperature exceeds 65 degrees Fahrenheit. Each month is summed to produce an annual total at the location of the building.
- `heating_degree_days`: heating degree day for a given day is the number of degrees where the daily average temperature falls under 65 degrees Fahrenheit. Each month is summed to produce an annual total at the location of the building.
- `precipitation_inches`: annual precipitation in inches at the location of the building
- `snowfall_inches`: annual snowfall in inches at the location of the building
- `snowdepth_inches`: annual snow depth in inches at the location of the building
- `avg_temp`: average temperature over a year at the location of the building
- `days_below_30F`: total number of days below 30 degrees Fahrenheit at the location of the building
- `days_below_20F`: total number of days below 20 degrees Fahrenheit at the location of the building
- `days_below_10F`: total number of days below 10 degrees Fahrenheit at the location of the building
- `days_below_0F`: total number of days below 0 degrees Fahrenheit at the location of the building
- `days_above_80F`: total number of days above 80 degrees Fahrenheit at the location of the building
- `days_above_90F`: total number of days above 90 degrees Fahrenheit at the location of the building
- `days_above_100F`: total number of days above 100 degrees Fahrenheit at the location of the building

- `days_above_110F`: total number of days above 110 degrees Fahrenheit at the location of the building
- `direction_max_wind_speed`: wind direction for maximum wind speed at the location of the building. Given in 360-degree compass point directions (e.g. 360 = north, 180 = south, etc.).
- `direction_peak_wind_speed`: wind direction for peak wind gust speed at the location of the building. Given in 360-degree compass point directions (e.g. 360 = north, 180 = south, etc.).
- `max_wind_speed`: maximum wind speed at the location of the building
- `days_with_fog`: number of days with fog at the location of the building

Target

- `site_eui`: Site Energy Usage Intensity is the amount of heat and electricity consumed by a building as reflected in utility bills

Your job

Highlight what you have learned in the machine learning course by building a sklearn pipeline to:
- correctly pre-process the data according to its category (we don't give the steps to take into account here as they can be found in the different courses),
- choose the right hyper-parameters for the models selected (we don't give the hyper-parameters to look for here, as they have already been covered in the course).

An initial study of the data (EDA) is obviously necessary in order to understand the nature of each of the features.

CAUTION: if you use concepts that you have not seen in class, you must be able to understand and explain them. There is no point in copying code found on the Internet or generated by ChatGPT without understanding it. We much prefer a less extensive but perfectly understood piece of work to an exhaustive piece of work consisting of applying a set of techniques without understanding how they work or why they are useful.

To be handed in: a written notebook (with comments) that you must present. This will have to justify your choices for the different stages of the pipeline.