# Project E18: KAGGLE ELECTRICITY CONSUMPTION

Johan Hollak, Meeri-Ly Muru, Anette Taivere
https://github.com/doktorjohan/IDS-project

# Task 2. Business understanding (400-800 words)

## Identifying your business goals

### Background

In the last few years due to various crises (COVID-19, climate change, labor shortages and the War in Ukraine) the cost of electricity has skyrocketed. Enefit is one of the largest energy companies in the Baltics and they would like to help their consumers on the way to reaching zero.  Both electricity cost and the environmental footprint could be drastically reduced by forecasting the consumption of a household.

### Business goals

- Find out how different weather aspects affects electricity consumption (air temperature, wind, rain etc)
- Find out how the hourly electricity price affects electricity consumption
- How time affects electricity consumption
- Predict the consumption value for each timestep
- Create an energy consumption prediction model for a single household
- Help people reach a more sustainable energy usage

### Business success criteria

- Gain enough insight of energy usage in households to build a prediction model
- The created prediction model has an accuracy of at least 95%
- The prediction model is helpful in helping Enefit consumers reaching zero

# Assessing your situation

### Inventory of resources

We have the electricity price and consumption for the period 2021-09-01 00:00 - 2022-08-24 23:00 for an individual household in Estonia and weather and the electricity price but not the consumption for the period 2022-08-25 00:00 - 2022-08-31 23:00.

### Requirements, assumptions, and constraints

Task is scheduled for completion 9.12.22.

Project presentation is on 15.12.22.

### Risks and contingencies

A power outage or loss of Internet connection. The contingency plan for this situation would be to continue the project in another building, given that it still has power or Internet (e.g. Delta Centre).

### Terminology

**time** - definition of example_id

**temp** - Air Temperature (°C)

**dwpt** - The dew point in °C

**rhum** - The relative humidity in percent (%)

**prcp** - The one hour precipitation total in mm

**snow** - The snow depth in mm

**wdir** - The wind direction in degrees (°)

**wspd** - The average wind speed in km/h

**wpgt** - The peak wind gust in km/h

**pres** - The sea-level air pressure in hPa

**coco** - The weather condition code

**el_price** - the electricity price in Estonia on that hour (€/kWh)

**consumption** - the electricity consumption (kWh)

### Costs and benefits

Since we're doing a Kaggle competition, we won't have any material costs. Working on the project will only cost time.

Benefits: Accurate household-level predictions could help people use energy more sustainably. Meaning, it would reduce electricity costs and it would also be good for the environment.

## Defining your data-mining goals

### Data-mining goals

- A model that predicts energy consumption for a single household
- A report describing the process of training the model
- A presentation describing the process of data mining

### Data-mining success criteria

- Model predicts with an accuracy of at least 95%
- Prediction output in acceptable format for kaggle

# Task 3. Data understanding (400-800 words)

## Gathering data

### Outline data requirements

**Time range:**
The training data is over a period of one calendar year, excluding the final week of august, which is considered to be the test data. The data points are gathered with an interval of one hour.

**Data relating to weather:**

- air temperature (degrees C, float)
- dew point (degrees C, float)
- relative humidity (percent, int)
- total precipitation of one hour in mm (int)
- snow depth in mm (float)
- wind direction in degrees (int)
- average wind speed in km/h (float)
- peak wind gust in km/h (float)
- sea-level air pressure in hPa (float)
- weather condition code (enumerated value, enum)

**Data relating to electricity:**

- electricity price in Estonia for the given hour (eur/kWh, float)
- current energy consumption for the given hour in kWh (float)

### Verify data availability

Data exists and is usable in kaggle.
https://www.kaggle.com/competitions/predict-electricity-consumption/data

### Define selection criteria

The data can be downloaded from
https://www.kaggle.com/competitions/predict-electricity-consumption/data

Kaggle presents us with the training and test data. The test data is split into two folds: a public and a private fold. There is no real need to acquire more data from other sources, although it is not forbidden and we will not exclude this option at this point.

## Describing data

We have 3 files: train.csv, test.csv and sample_submission.csv.
Train.csv has data of the training set which includes the weather, electricity price and the electricity consumption for the period 2021-09-01 00:00 - 2022-08-24 23:00 for an individual household in Estonia. The train.csv data size is 736.98 kB.

Test.csv has the test set data which includes the weather and the electricity price but not the consumption for the period 2022-08-25 00:00 - 2022-08-31 23:00 (the next seven days after the last timestep in the training data).  The test.csv data size is 13.96 kB.

Our Kaggle task is to predict the consumption of this household for these next 7 days.
Sample_submission.csv is a sample submission file in the correct format. The submission file needs to include only two columns: time & consumption. The sample_submission.csv is 5.37 kB.

## Exploring data

**Problems:**

- Snow and precipitation is not always present, a lot of empty values
- Weather condition code must be encoded using one-hot encoding
- Problems regarding data types in columns, columns containing float and int types

**Value ranges (min, max):**

- time ()
- temp (-26.1, 31.4)
- dwpt (-28.7, 20.9)
- rhum (0, 100)
- prcp (0, 7.9)
- snow (0, 220)
- wdir (0, 360), wind blows in every direction
- wspd (0, 31.7)
- wpgt (2.9, 63)
- pres (962.6, 1047.5)
- coco (1, 25)
- el_price (0.000070, 4)
- consumption (0, 10.38)

## Verifying data quality

Apart from the minor problems already listed above, the data is of high quality. This project faces no serious data quality issues, such as missing data or non-accessible data.

# Task 4. Planning your project (100-300 words)

**Tasks:**

1. Set up an environment and access to data. One team member for 1 hour.
2. Filtering and cleaning the data. Each team member for 1 hour.
3. Find patterns and trends then draw out the relevant results. Each team member for 5 hours.
4. Experiment with different approaches and adding features. Each team member for 5 hours.
5. Modeling. Train numerous models to define which one of them provides the most accurate predictions. Each team member for 10 hours.
6. Finishing touches and final submission to Kaggle. Each team member for 2 hours.
7. Poster session preparation - each team member for 4 hours
8. Meetings - each team member for 3 hours

**Tools:**

- Python
- Jupyter Notebook
- Pandas
- Excel
- Anaconda3
- tensorflow/keras/scikit-learn
- Github for hosting the project repository