

Your workshop companion

Machine Learning: From Data to Decisions

Third-year engineering / Basic Python / One teaching day

The case study

Campus Energy Intelligence predicts end-of-day energy and flags high-demand scenarios. All 600 rows are fictional independent scenarios. No result measures any real campus performance.

What you will keep

Six executed notebooks, a comparison against a simple baseline, a confusion matrix, a leakage investigation and your model card. Use a team alias in files. Keep your explanations in your own words.

The day

09:00 introduction; 09:30 workflow; 10:00 data; 10:45 break; 11:00 regression; 12:00 classification; 12:45 lunch; 13:30 evaluation; 14:15 trees; 15:00 break; 15:15 challenge; 16:15 review; 16:40 exit.

Python bridge

A DataFrame is a table. `df[columns]` selects inputs. `model.fit(X,y)` learns from examples. `model.predict(X)` returns outputs. Run cells in order and read the final line of an error first.

The experiment contract

Before a model

Define a decision, who uses it, when it is made and which inputs exist then. A predictor learns statistical associations; it does not establish causation or automatically choose a responsible action.

Two targets

Regression predicts energy_kwh. Classification predicts high_demand, which is 1 when observed energy is at least 125 kWh. This threshold is a classroom definition.

Three separate roles

Training: 360 rows to fit imputation, scaling and models. Validation: 120 rows to compare depth or probability thresholds. Test: 120 rows kept until decisions are frozen. Never tune against final-test results.

Why this split

The generator samples independent fictional scenarios. Real campus days may share buildings and time patterns, so random row splitting could overstate generalisation. Use building groups or future periods for that different question.

Safe preprocessing

SimpleImputer learns training medians. StandardScaler learns training means and scales. Pipeline keeps these fitted steps attached to the model. Calling fit on validation or test leaks information.

01. Data preparation

A data inspection routine

Read the data card and CSV. Check shape, units, missing values and allowed features. Exclude `scenario_id` from learning. Do not silently fill every missing value with zero. Zero occupancy is a meaningful value and is not the same as an unknown measurement.

Notebook objective

Inspect the CSV and identify which columns can exist at prediction time.

Work to submit to your own folder

Draw an input/target table. Explain why the ID is excluded. Count missing values on training and describe how validation uses training medians.

Partner checkpoint

Explain the units, data subset and decision behind the number. Swap driver/navigator roles. Save your notebook and one observation before advancing.

Where it fits

10:00–10:45; use the matching numbered notebook. The workspace includes an offline ZIP.

02. Linear regression

Numerical prediction

Linear regression fits an intercept plus weighted inputs by minimising squared residuals. Our reported MAE is the average absolute error in kWh. Compare with DummyRegressor trained on the training mean. A model needs to beat a meaningful simple baseline on unseen rows.

Notebook objective

Predict numerical energy and compare against a training-mean baseline.

Work to submit to your own folder

Calculate MAE by hand for actual [100,120,140] and predicted [110,115,135]. Try occupancy-only regression using the same validation rows. Record whether all four inputs improve MAE.

Partner checkpoint

Explain the units, data subset and decision behind the number. Swap driver/navigator roles. Save your notebook and one observation before advancing.

Where it fits

11:00–12:00; use the matching numbered notebook. The workspace includes an offline ZIP.

03. Logistic regression

Probability and class

Logistic regression produces a score mapped into 0–1. A 0.5 decision threshold is a convention. It differs from the 125 kWh rule that defines observed high demand. Probabilities can be miscalibrated and performance changes across populations.

Notebook objective

Classify high demand with a probability model.

Work to submit to your own folder

Explain the difference between 125 kWh label definition and 0.5 probability threshold. Record a false negative and false positive scenario.

Partner checkpoint

Explain the units, data subset and decision behind the number. Swap driver/navigator roles. Save your notebook and one observation before advancing.

Where it fits

12:00–12:45; use the matching numbered notebook. The workspace includes an offline ZIP.

04. Evaluation and thresholds

Reading a confusion matrix

Rows are actual labels, columns predicted, labels=[0,1]: first row TN,FP; second FN,TP.

Accuracy=(TP+TN)/N; precision=TP/(TP+FP); recall=TP/(TP+FN). The notebook uses zero_division=0 so undefined ratios report zero; interpret the counts too. Choose thresholds on validation, using an explicit cost assumption.

Notebook objective

Choose a decision threshold using validation evidence.

Work to submit to your own folder

For TP=18, FP=6, FN=2, TN=74 compute accuracy, precision and recall. Change missed-event cost from 5 to 2 using validation only.

Partner checkpoint

Explain the units, data subset and decision behind the number. Swap driver/navigator roles. Save your notebook and one observation before advancing.

Where it fits

13:30–14:15; use the matching numbered notebook. The workspace includes an offline ZIP.

05. Trees and overfitting

Complexity and overfitting

A decision tree asks successive feature questions. Increasing `max_depth` can improve training fit while hurting unseen performance. Compare training and validation F1. One fixed split gives an exercise, not robust research evidence; later work can use cross-validation and confidence intervals.

Notebook objective

Compare tree complexity and the training/validation gap.

Work to submit to your own folder

Choose a tree depth from the validation table and justify it. Explain why maximum training F1 is insufficient. Sketch the first two decision rules.

Partner checkpoint

Explain the units, data subset and decision behind the number. Swap driver/navigator roles. Save your notebook and one observation before advancing.

Where it fits

14:15–15:00; use the matching numbered notebook. The workspace includes an offline ZIP.

06. Leakage investigation and final test

Evidence and leakage

end_of_day_meter_kwh duplicates the observed outcome and is only available after the prediction time. Its near-perfect score is invalid. Remove unavailable inputs, train a correct pipeline, select on validation, record the choice and then report the untouched test once.

Notebook objective

Audit a suspicious result, freeze decisions and report final evidence.

Work to submit to your own folder

Write a model card with decision, inputs, exclusions, split, validation choice, test metrics, failure and limitation. Present why near-zero MAE was invalid.

Partner checkpoint

Explain the units, data subset and decision behind the number. Swap driver/navigator roles. Save your notebook and one observation before advancing.

Where it fits

15:15–16:15; use the matching numbered notebook. The workspace includes an offline ZIP.

After the workshop

Seven-day extension

Day1 reproduce the guided result. Day2 write the model card. Day3 compare one new model on validation. Day4 inspect a failure. Day5 design a building/time split for real data without collecting it. Day6 explain findings to a peer. Day7 submit a short portfolio note locally or through the college-approved channel.

Responsible use

Do not send personal student records, campus credentials or confidential data to online tools. If AI helps with code, inspect the changes and explain the result yourself. Never load model pickle files from untrusted sources.

Further reading

https://scikit-learn.org/stable/common_pitfalls.html

https://scikit-learn.org/stable/modules/classification_threshold.html

https://scikit-learn.org/stable/modules/cross_validation.html

Participation

Participation certificates record attendance verified by the organiser. They do not certify academic credit, professional competence or college accreditation. No certificate is issued by this workbook.