

Machine Learning: From Data to Decisions

Workshop 01
Third-year engineering
Programme faculty: Akash Chatake



The campus-energy decision

Before the day begins, estimate energy use.
Flag scenarios that may exceed 125 kWh.
All examples today are synthetic.

Today's working files

Six independent notebooks

One CSV and a model card

A practical record and individual exit response

The day at a glance

Morning: data, regression and classification

Afternoon: evaluation, trees and leakage

Breaks: 10:45, 12:45 and 15:00

Close: 17:00

Rules and learned models

A rule executes a decision we specify.

An ML model estimates a pattern from examples.

A generative model produces content.

Every output still needs a suitable check.

Regression and classification

Regression: how many kWh?

Classification: high demand or normal?

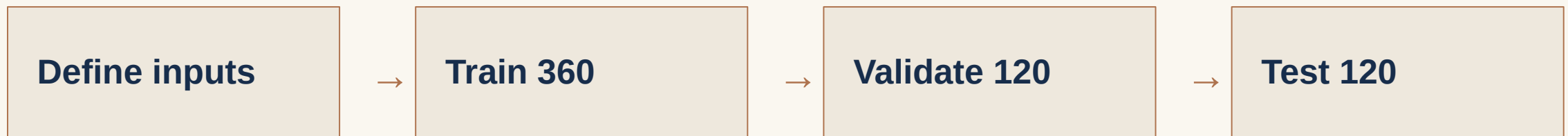
The choice of target determines the task.

The experiment flow

Training fits the model.

Validation chooses a setting.

The held-out test reports the final result.



Prediction-time availability

Available inputs: temperature, occupancy, operating hours and area.

Observed energy arrives after the day ends.

An identifier is not a measurement.

Practical 01: Data preparation

Open notebook 01. Run the setup cell.
Inspect training rows and missing values.
Write an input/target table.
Explain the split to your partner.

Missingness and preprocessing

Unknown occupancy differs from zero occupancy.
Training medians fill missing values.
Validation reuses those medians.
Pipeline carries fitted steps into prediction.

Practical 02: Numerical prediction

Fit a training-mean baseline.

Fit linear regression on the same training rows.

Compare validation MAE in kWh.

Inspect the prediction plot.

Mean absolute error

Actual: 100, 120, 140 kWh

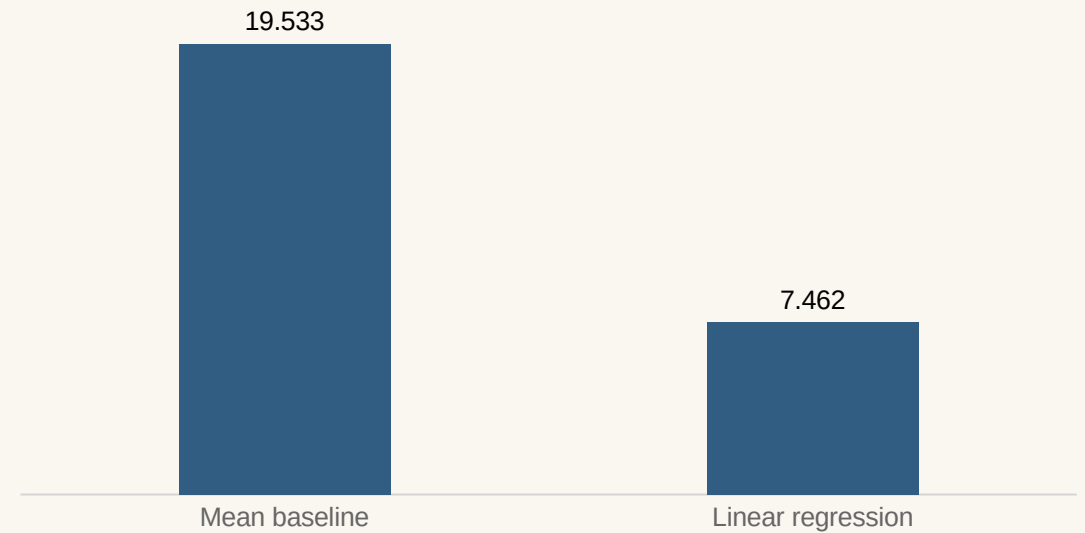
Predicted: 110, 115, 135 kWh

MAE = mean of the absolute errors

Calculate it before running the model.

Validation comparison

Same 120 validation scenarios
Same four allowed inputs
Lower MAE means smaller average error here.



Practical 03: High-demand classification

Label: observed energy at least 125 kWh.

Model: logistic regression.

Prediction: probability followed by a threshold.

Compare a majority-class baseline.

Probability and action

A probability is a model estimate.

A threshold converts it into an alert.

The choice depends on missed events and false alarms.

Practical 04: Evaluation

Read the confusion counts.
Calculate accuracy, precision and recall.
Compare thresholds using validation only.

Confusion matrix

Actual rows and predicted columns

TN: correct normal FP: false alarm

FN: missed high demand TP: correct alert

A threshold needs a stated assumption

Classroom cost = $5 \times$ missed events + false alarms

Compare 0.3, 0.4, 0.5, 0.6 and 0.7.

Freeze the choice before looking at test data.

Practical 05: Decision trees

Compare depths 1, 2, 3, 5 and unrestricted.
Record training and validation F1.
Explain one training/validation gap.

Complexity and unseen data

A model can memorise noise.
Training fit is insufficient evidence.
Validation supports the complexity choice.
One split still carries sampling uncertainty.

Practical 06: A suspicious result

An extra column produces an excellent score.
Audit when each input becomes available.
Repair the experiment and freeze the choice.

Final held-out evidence

Record the frozen features and settings.

Run the final test once.

Keep the result even when it disappoints.

Describe a failure and the limits of synthetic data.

The model card

Decision and input availability

Data, split and baseline

Validation choice and final metrics

Failure, limitation and next experiment

Individual exit response

Feature or target?

Why train-only imputation?

What is a false negative?

When does the test set open?

Why can a perfect score mislead?

After today

Reproduce the experiment.
Document one improvement and one failure.
Keep your notebook and model card.
Use the optional podcasts for revision.

Workshop resources

learn.mindforgeai.co.in/workshops/workshop-01/index.html

edusphere@chatakeinnoworks.com

Jigyasa by MindForgeAI