SH Dr. Soufiane Hourri

drsoufiane.com

DUT · SEMESTER 3 · COMPUTER SCIENCE

Machine Learning

Practice by hand · Describing and preparing data (Sessions 01 and 02)

Dr. Soufiane Hourri, PhD, HDR

Associate Professor & AI Consultant

École Supérieure de Technologie de Safi · Cadi Ayyad University

2026-2027

PRACTICE · OVERVIEW

One hour of practice, by hand

01 Screen time: center and spread · 12 min

02 Sleep and score: correlation · 10 min

03 Splitting: proportions · 8 min

04 Missing incomes: mean or median? · 13 min

05 Encoding and scaling · 15 min

HOW WE WORK

In pairs, with paper and a calculator.

Every result with its unit and one sentence of interpretation.

PRACTICE · REMINDERS

Formula card

Quantity	Formula	Session
Mean · median	$\bar{x} = \Sigma\, x_i / n$ · middle value of the sorted data	01
Standard deviation	$s = \sqrt{\Sigma\,(x_i - \bar{x})^2 / (n - 1)}$	01
Quartiles, outliers	IQR = Q3 − Q1; fences Q1 − 1.5 × IQR and Q3 + 1.5 × IQR	01
z-score	$z = (x - \bar{x}) / s$	01
Correlation	$r = \Sigma\,(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\Sigma\,(x_i - \bar{x})^2 \times \Sigma\,(y_i - \bar{y})^2}$	01
Min-max scaling	$x' = (x - \min) / (\max - \min)$	02

Every preparation value (mean, median, s, min, max, categories) is learned on the training set only.

Exercise 1 · Screen time

DATA (EXAMPLE)

Hours of screen time per day for nine students: 5, 3, 12, 4, 5, 2, 6, 3, 5

TIME

12 min

QUESTIONS

1. Mean, median and mode
2. Range and standard deviation s (with n − 1)
3. Q1, Q3 and IQR. Is any value an outlier?
4. z-scores of 12 h and of 2 h. Interpret
5. Remove 12 h: how do the mean and the median change?

ANSWER FORMAT

Each result with its unit; one sentence for question 5.

Exercise 2 · Sleep and test score

DATA (EXAMPLE)

Five students: hours of sleep before an exam and score out of 20: (5, 8), (6, 10), (7, 9), (8, 13), (9, 15)

TIME

10 min

QUESTIONS

1. Compute the means and the standard deviations of both variables
2. Compute the correlation r (table of deviations and products)
3. Interpret the sign and the strength of r
4. Can we conclude that more sleep causes a better score? Why?

TIP

Build the table: $x_i - \bar{x}$, $y_i - \bar{y}$, their product, and both squares.

Exercise 3 · Splitting a dataset

DATA (EXAMPLE)

250 students, 40 of whom failed. A model will predict failure; 20% of the data goes to the test set.

TIME

8 min

QUESTIONS

1. What proportion of students failed?
2. How many students in the test set? With a stratified split, how many failed students in each set?
3. A random split put only 3 failed students in the test set. What proportion? Why is it a problem?
4. The file is sorted by grade. Why not take the last 50 rows as the test set?

ANSWER FORMAT

Counts and percentages; one sentence for questions 3 and 4.

Exercise 4 · Missing incomes: the data

Household	Monthly income (thousand MAD)	City
H1	4	Safi
H2	5	Safi
H3	5	Marrakech
H4	6	Safi
H5	?	Agadir
H6	7	?
H7	?	Marrakech
H8	21	Safi

Training set, example data. ? = missing value.

Exercise 4 · Mean or median?

GOAL

Choose how to fill the gaps, and measure what filling does to the statistics.

TIME

13 min

QUESTIONS

1. What share of incomes is missing?
2. Mean, median and s of the observed incomes
3. Fill with the mean, then with the median: new mean and s?
4. Which fill value do you choose? Why?
5. A test household has no income. Which value fills it?
6. How do you fill the missing city of H6?

ANSWER FORMAT

Numbers in thousand MAD; one sentence for questions 4 to 6.

Exercise 5 · Encoding and scaling

DATA (EXAMPLE)

Training cities: Safi, Marrakech, Safi, Agadir. Training ages: 18, 19, 20, 21, 22.

TIME

15 min

QUESTIONS

1. One-hot encode the four cities. How is Essaouira encoded?
2. Encode the Bac mention. Why not number the cities 1, 2, 3?
3. z-scores and min-max values of the five ages
4. Scale a test student aged 24. Which mean and s?
5. A (19 y, 5,000 MAD), B (21 y, 5,100 MAD), s of income 1,000: distance raw, then scaled?

TIP

Distance = $\sqrt{(\Delta age^2 + \Delta income^2)}$.
Scaled: divide each difference by its s.

To finish: one question

QUESTION

You standardise ages before training a model. Where do the mean and s come from?

A

From the whole dataset, before splitting

B

From the training set only, reused on the test set

C

From the training set for training, from the test set for testing

D

From the training and test sets together

SH

Well done · Questions?

Keep your sheets: these calculations come back in the exam.

Dr. Soufiane Hourri, PhD, HDR

drsoufiane.com