Grade 12 Computing

Unit 6 · Building and Judging AI Systems

ANSWER KEY · FOR TEACHERS

Full marks 35

Give the marks for the points in the marking scheme; an answer in other words that says the same is right.

1 Key words

1. supervised learning
2. clustering
3. reward
4. Gini impurity
5. precision
6. recall
7. hyperparameter
8. prompt injection

Each carries 0.5 mark. **K · 4**

2 Label the diagram

1. asks for a tool
2. its result
3. Limits set by its builders
4. Tools: Python functions

Each carries 0.5 mark. **U · 2**

3 Multiple choice

i. **b.** Every district in that pile has the same belt. **U · 1**

A Gini impurity of 0 means two cards drawn from the pile always have the same label: the pile is pure, here all Terai. The tree stops growing there and makes it a leaf. The other pile, still mixed, is split again.

ii. **b.** about 9% **A · 1**

Precision = TP ÷ (TP + FP) = 9 ÷ (9 + 90) ≈ 9%: of the 99 people it says have the disease, only 9 do. Its recall is 9 ÷ 10 = 90%, which is good. When the thing looked for is rare, even a small share of false alarms among the many healthy people swamps the true finds; that is why a positive screening test is followed by a second, better test.

iii. **c.** the number of clusters, k, given to k-means **U · 1**

k is chosen by the person, before k-means runs, and controls how it works. The cut-off, the weights and the final centers are all learned from the data: they are parameters.

iv. **c.** Making the payment tool require a person's approval, and allowing payments only to accounts already on an approved list. **A · 1**

A model may be talked out of its instructions by the email itself, and newer models are still fooled. A check in the payment tool (a person's approval, a list of allowed accounts) works whatever the model is persuaded of. Finding out at the end of the month is too late: the money is gone.

v. **b.** 0.25 **A · 1**

(30 × 0.2 + 10 × 0.4) ÷ 40 = (6 + 4) ÷ 40 = 0.25. The bigger pile counts for more, so the answer is nearer 0.2 than the plain average, 0.30.

4 What does it print?

MARKING SCHEME **U · 3**

- ✓ `0.0`: a pile of one label is not mixed at all ($1 - 1^2$).
- ✓ `0.5`: two labels, half each: $1 - (0.25 + 0.25)$.
- ✓ `0.56`: $1 - (0.6^2 + 0.2^2 + 0.2^2) = 1 - 0.44$.

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

5 Trace k-means

ANSWER A · 4

Round	`groups`	`centers` **after moving**
1	[0, 1, 1, 1, 1, 1]	[1.0, 7.4]
2	[0, 0, 0, 1, 1, 1]	[2.33, 10.33]

In round 3 every point joins the same center as in round 2, so `new == groups` and the loop breaks before printing: nothing changed, so k-means has finished.

MARKING SCHEME

- ✓ Round 1: only 1 is nearer center 1; the other five join center 2: `[0, 1, 1, 1, 1, 1]`.
- ✓ Round 1's centers: 1.0, and the mean of 2, 4, 9, 10 and 12, 7.4.
- ✓ Round 2: 2 and 4 are now nearer 1.0 than 7.4: `[0, 0, 0, 1, 1, 1]`, centers 2.33 and 10.33.
- ✓ It stops in round 3, when no point changes group.

6 Precision and recall

ANSWER A · 4

TP = 28, FN = 30 − 28 = 2, FP = 174, TN = 470 − 174 = 296.

- Accuracy = (28 + 296) ÷ 500 = 324 ÷ 500 = **64.8%**
- Precision = 28 ÷ (28 + 174) = 28 ÷ 202 = **13.9%**
- Recall = 28 ÷ (28 + 2) = 28 ÷ 30 = **93.3%**
- F1 = 2 × 0.1386 × 0.9333 ÷ (0.1386 + 0.9333) = 0.2587 ÷ 1.0719 = **24.1%**

Neither is simply better: at 0.20 it catches nearly every scam, but only about 1 flag in 7 is real, so people may learn to ignore the warning. Which is right depends on what each mistake costs: for a warning, a missed scam costs more than a false alarm, so a threshold lower than 0.50 may be right; for a filter that deletes messages, it isn't.

MARKING SCHEME

- ✓ The four counts: TP 28, FN 2, FP 174, TN 296.
- ✓ Accuracy 64.8% and recall 93.3%.
- ✓ Precision 13.9% and F1 24.1% (24.2% from the rounded figures).
- ✓ A judgment from the costs of the two mistakes: more scams caught, but most flags false.

7 Write a program

ANSWER

```
APPROVED = ["9841000001", "9841000002"]    # the family's own phones
sent = []


def send_sms(number, text):
    """Sends a text from the family's phone, with its limits inside the tool."""
    if number not in APPROVED:
        return {"error": "not an approved number"}
    if len(sent) >= 3:
        return {"error": "no more than 3 messages in a run"}
    sent.append(number)
    print("SMS to", number + ":", text)
    return "sent"


print(send_sms("9800000000", "Your tickets: T1001 T1002 T1003"))
for i in range(4):
    print(send_sms("9841000001", "Ticket T1001, bus B3, Saturday 7:30"))
```


Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

A · 4

```
OUTPUT
{'error': 'not an approved number'}
SMS to 9841000001: Ticket T1001, bus B3, Saturday 7:30
sent
SMS to 9841000001: Ticket T1001, bus B3, Saturday 7:30
sent
SMS to 9841000001: Ticket T1001, bus B3, Saturday 7:30
sent
{'error': 'no more than 3 messages in a run'}
```

Whatever text the model has read, the code refuses a stranger's number and a fourth message.

MARKING SCHEME

- ✓ A list of approved numbers, and a check that refuses any other number.
- ✓ A count of the messages sent (a list or a counter kept outside the function).
- ✓ A check that refuses once 3 have been sent.
- ✓ Both refusals return an error dictionary before anything is sent; a message that passes is printed and `"sent"` returned.

8 Short answer

MARKING SCHEME — U · 3

- ✓ a. A plain program: the rules are written down exactly, and a wrong grade would harm a student.
- ✓ b. A neural network: the data is photos; it should advise, not decide, and be tested on photos from Nepal's farms.
- ✓ c. A decision tree (or another small, readable model): a table with labels, and applicants must be told why; checked for each group, with a person deciding.

9 Long answer

ANSWER — U · 4

How: the model learned from the messages the company had, which were mostly in English, so it had seen few Nepali scams and learned few of their signs. Its overall recall hid the gap, because English messages were the majority.

Better data: collect and label many more Nepali messages (scams and normal), with permission, and retrain. This fixes the cause: in the unit's example the retrained model caught 13 of 15 scams in each language with few false alarms. It costs time and money.

A lower threshold for Nepali messages: quick, but the model's Nepali scores are poor, so catching more scams brings many more false alarms (at 0.30, 9 of 15 scams caught but 39 normal messages flagged, so only 19% of its Nepali flags are real). Treating groups differently also needs a clear reason. It may help for a while, but better data is the real fix.

MARKING SCHEME

- ✓ It learned mostly from English messages, so it saw few Nepali scams and learned few of their signs.
- ✓ The overall recall hid the gap, because English messages were the majority.
- ✓ Better data: collect and label more Nepali messages, with permission, and retrain; this fixes the cause, at a cost of time and money.
- ✓ A lower threshold for Nepali: quick, but it brings many false alarms (at 0.30, only 19% of Nepali flags real) and treats groups differently; better data is the real fix.

10 Think

ANSWER

Not necessarily. A written rule is explainable, which makes its unfairness easier to see and to challenge, but it can still be unfair, because it carries the values of the person who

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking

wrote it. For example, a cooperative rule "loans only to applicants with land in their own name" is exact and readable, but shuts out most women where land is usually registered in men's names. A model learned from past decisions can repeat past unfairness too. Either way, a rule or a model must be checked for whom it fails, for each group it affects, and the values in its design must be stated and questioned.

MARKING SCHEME

- ✓ Not necessarily: a readable rule makes unfairness easier to see, but it still carries its writer's values, as a model carries its data's.
- ✓ An example, such as the cooperative's "land in your own name" rule, which shuts out most women; either must be checked for whom it fails.

11 Activity: Run the agent, then limit it

(not marked)

WHAT STUDENTS SHOULD FIND

Cheapest bus: it booked the Night Rider, arriving at 4:30 am: what Sita said, not what she meant; no limit stops it, only asking the user would. **Before 3 pm:** nothing went wrong: two tool calls, no payment. **The offer page:** it obeyed the instruction in the page (a prompt injection), booked three seats and texted the tickets to a stranger; "Ask me" lets the user refuse the extra bookings and the message. **The full bus:** it kept retrying a full bus until the limit on tool calls stopped it, sooner at 6. Good students notice that every harm was done through a tool, so the limits that held were in the tools and the program, not in the model.

Marks by level	Knowledge	Understanding	Application	Higher ability	**Total**
Marks	4	14	15	2	**35**

Science Playground, Katunje, Bhaktapur
Book a class visit: https://science-playground.com/booking