2 points - no penalty for a wrong answer

A KNN classifier is trained on the training set in the table below, which only contains categorical features.

	width	weight	speed	Class
x1	big	light	fast	A
x2	big	heavy	fast	B
x3	small	heavy	slow	B
x4	big	heavy	slow	A
x5	small	light	slow	A

With categorical features, the distance between two samples *a* and *b* can be computed as the number of features with different values:

$$dist(a, b) = \sum_i \ell(a_i, b_i)$$

Where:

$$\ell(m, n) = \begin{cases} 0 & if\ m = n \\ 1 & otherwise \end{cases}$$

For each neighbor x_i of x the vote is weighted as follows:
$weight(x, x_i) = \frac{1}{dist(x, x_i)+0.5}$

Given the test point below, write in the answer box:

1. The list of the 3 neighbors of t1
2. The class assigned to t1 with K = 3

	width	weight	speed	Class
t1	big	heavy	fast	?

Use the following notation:
neighbors={list of neighbors for t1}
class=class assigned

Example:
neighbors={point1 point2 point3}
class=A

Domanda 2

Risposta non data

Punteggio max.: 1,00

1 point - no penalty for a wrong answer

The z-score normalization is applied to a feature x by removing the mean μ and dividing by its standard deviation σ. The resulting feature x' is:

$$x' = \frac{x-\mu}{\sigma}$$

Consider the regression task described by the training set and the test set below, in which **the features have not been normalized**. The mean and the standard deviation (std) for each column are included at the bottom of the tables for your convenience.

The dataset contains **three input features** (freq, rpm, power) and an **output variable** (Y).

The input features have been normalized with z-score (not shown in the tables) before training a regression model to predict the value of Y.

You are required to perform the normalization on the test point t1 (see test set table): write in the answer box the normalized row corresponding to the test point t1, after applying the proper z-score normalization on the features.

Training set

	freq	rpm	power	Y
x1	10	20	0	0
x2	5	0	2	10
x3	5	0	0	10
x4	10	20	2	40
mean	7.5	10	1	15
std	2.5	10	1	15

Test set

	freq	rpm	power	Y
t1	5.5	1	4	3
t2	0.5	6	8	6
mean	3	3.5	6	4.5
std	2.5	2.5	2	1.5

Use the following notation:
t1=[freq, rpm, power, Y]

Example:
t1=[1, 2, 3, 4]

Domanda 3

Risposta non data

Punteggio max.: 2,00

2 points - no penalty for a wrong answer

The Gini index of a node is computed as follows:

$$gini(node) = 1 - \sum_j P(j|t)^2$$

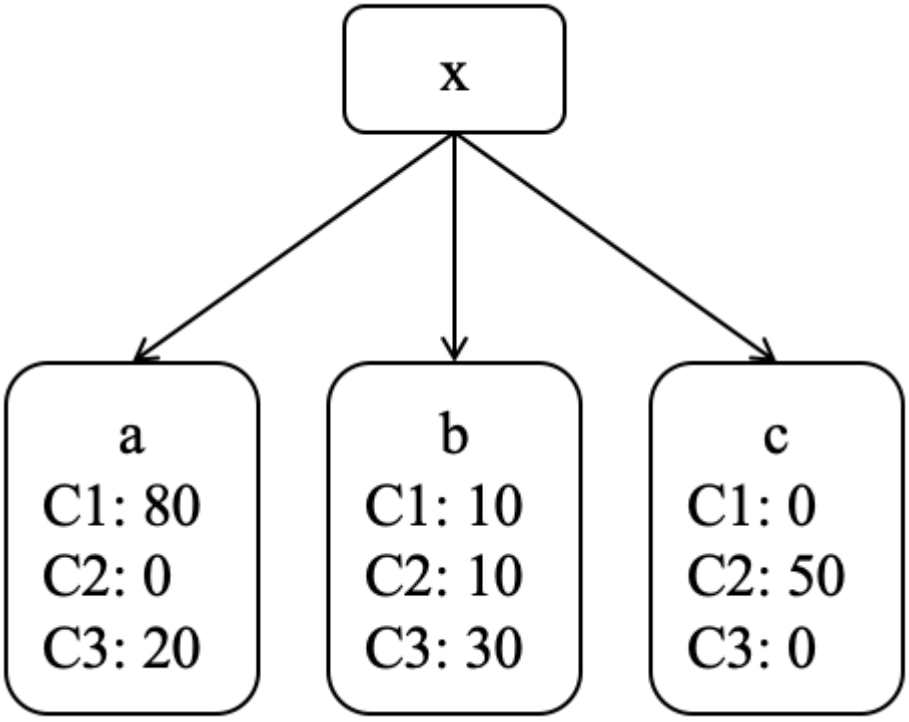
where $P(j|t)$ is the relative frequency of class j at node t

The Gini index of a *split* with parent *p* and children c_i is computed as follows:

$$gini(split) = \sum_i \frac{n_i}{n} gini(c_i)$$

where n_i is the number of records at child c_i and *n* is the number of records in *p*.

In the figure below it is shown a split *x*, with three children (*a, b, c*). For each child you are given the number of elements belonging to each of three classes (C1, C2, C3).



Write in the box below:
1. the Gini index of child a
2. the Gini index of the split x

Use the following notation:
gini(a)=value
gini(x)=value

Example
gini(a)=1/2
gini(x)=1/3

Domanda 4

Risposta non data

Punteggio max.: 2,00

2 points - no penalty for a wrong answer

Given the **similarity matrix** in the figure, apply agglomerative hierarchical clustering with **single (MIN) linkage**. The similarity metric ranges from 0 (more distant) to 1 (less distant).

	a	b	c	d	e
a	1	0.4	0.4	0.8	0.5
b	0.4	1	0.3	0.3	0.4
c	0.4	0.3	1	0.4	0.9
d	0.8	0.3	0.4	1	0.5
e	0.5	0.4	0.9	0.5	1

Let K be the number of obtained clusters. Write in the answer box:

1. The clusters obtained with K=3
2. The clusters obtained with K=2

Use the following notation:
K=3 {points of the first cluster} {points of the second cluster} {points of the third cluster}
K=2 {points of the first cluster} {points of the second cluster}

Example:
K=3 {point1 point2} {point3 point4} {point5}
K=2 {point1 point2} {point 3 point4 point5}

Domanda 5

Risposta non data

Punteggio max.: 1,00

1 point - -15% penalty for a wrong answer

Consider the following numpy array:

```
x = np.array([
              [2, 4,  6,  8],
              [3, 6,  9, 12],
              [4, 8, 12, 16]
])
```

The following indexing techniques are applied to obtain a and b.

```
a = x[1:, ::-1]
b = x[[False, True, True], :0:-1]
```

Which of the following assertions is correct?

- (a) The value of a is [[3, 6, 9, 12], [4, 8, 12, 16]]. The variables a and b contain different values.
- (b) The value of a is [[12, 9, 6, 3], [16, 12, 8, 4]]. The variables a and b contain the same values.
- (c) The value of a is [[3, 6, 9, 12], [4, 8, 12, 16]]. The variables a and b contain the same values.
- (d) The value of a is [[12, 9, 6, 3], [16, 12, 8, 4]]. The variables a and b contain different values.
- (e) None of the other answers is correct.
- (f) The script will provide an error because b cannot be computed.

Domanda 6

Risposta non data

Punteggio max.: 2,50

2.5 points - no penalty for a wrong answer

For each data point p_i

let $intra(p_i)$ be the average dissimilarity of p_i with all other points within the same cluster
let $inter(p_i)$ be the lowest average dissimilarity of p_i to any other cluster, of which p_i is not a member

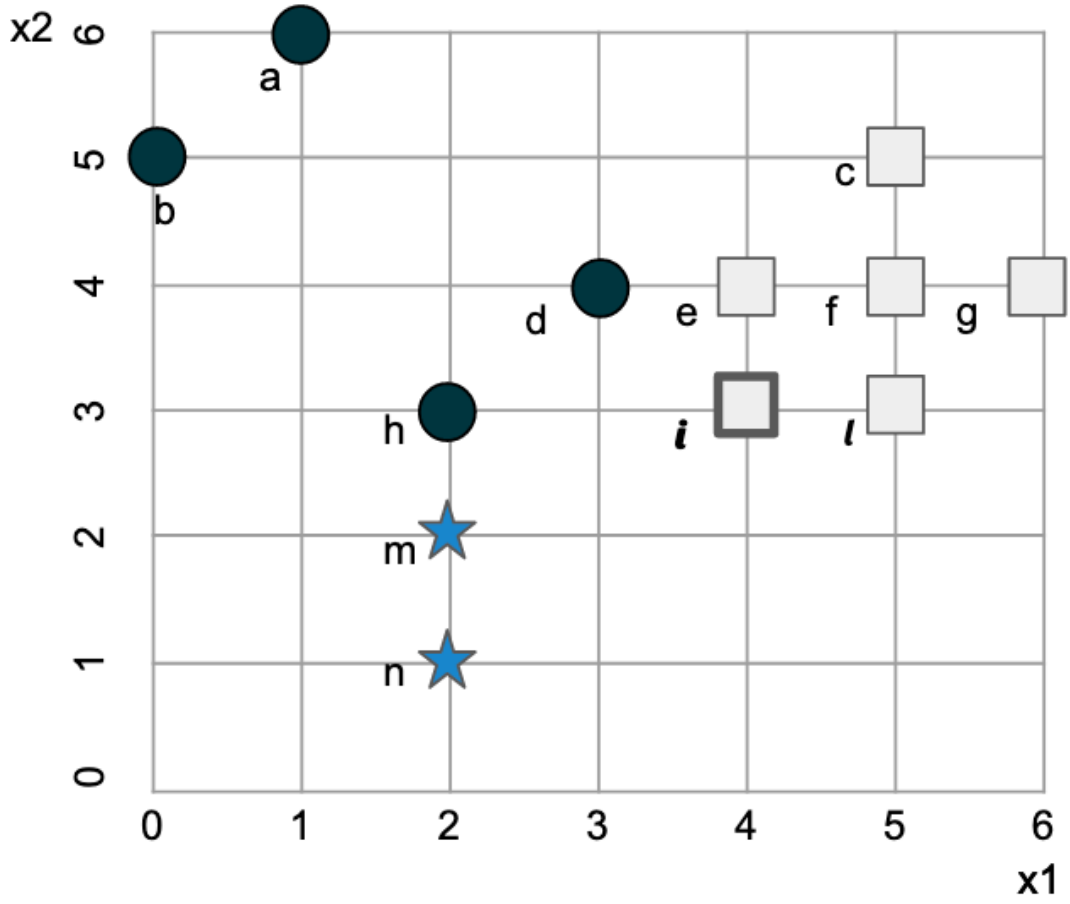
The Silhouette score of point pi is defined as:

$$silh(p_i) = \frac{inter(p_i) - intra(p_i)}{max(inter(p_i), intra(p_i))}$$

For two n-dimensional points $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n)$, the Manhattan distance is defined as follows:

$$dist(a, b) = \sum_i^n |a_i - b_i|$$

In the figure below, three different clusters are represented with the following shapes: circle, star, square. Use the Manhattan distance as the dissimilarity metric.



Consider point *i* belonging to the square cluster. Write in the answer box:

- intra(i)
- inter(i)
- silh(i)

Use the following notation:
intra(i)=value
inter(i)=value
silh(i)=value

Example:
intra(i)=2.5
inter(i)=5

silh(i)=2.5/5

Domanda 7

Risposta non data

Punteggio max.: 1,50

1.5 points - -15% penalty for a wrong answer

An itemset I is closed if none of its immediate supersets has the same support as I.

An itemset I is represented by a collection of literals (e.g. abc) and a number representing its support count (e.g. abc: 12).

After the analysis of a transactional dataset, the list of frequent itemsets found (with support counts) is the following:

length=1 a:150 b:160 c:180 d:150 e:150
length=2 ab:140 ac:120 ad:130 ae:140
length=3 abc:120 acd:110 abe:130
length=4 abce:120

Which of the following statements is true?

- (a) The only non-closed itemsets are **ac**, **ad**, **abc**, **abce**
- (b) The only non-closed itemsets are **ac**, **ad**
- (c) None of the other answers is correct
- (d) The only non-closed itemsets are **ac**, **ad**
- (e) The only closed itemsets are **ac**, **abc**
- (f) The only non-closed itemsets are **ac**, **abc**

Domanda 8

Risposta non data

Punteggio max.: 1,00

1 point - -15% penalty for a wrong answer

- Precision(C) is the fraction of correct predictions among the samples predicted with class C
- Recall(C) is the fraction of correct predictions among the samples with actual class C

After training two models (**M1**, **M2**), we obtain the confusion matrices shown below. Consider class **b** only. Which of the following statements is true?

predicted class

true class / **M1**	a	b	c	d
a	15	10	1	0
b	0	50	5	5
c	5	20	70	0
d	5	10	0	60

predicted class

true class / **M2**	a	b	c	d
a	14	10	0	2
b	5	40	5	10
c	5	5	80	5
d	10	5	10	50

(a) None of the other answers is correct

- (b) M1 has lower recall than M2 and lower precision.
- (c) M1 has the same recall as M2 and lower precision.
- (d) M1 has higher recall than M2 and lower precision.
- (e) M1 has higher recall than M2 and higher precision.
- (f) M1 has lower recall than M2 and higher precision.

Domanda 9

Risposta non data

Punteggio max.: 1,50

1.5 points - -15% penalty for a wrong answer

Consider the following Python operations:

```
l1 = [1,2,3]
l2 = [2,4,6]
d = {'mon':l1, 'tue':l2}
res = d.values()
l1.append(4)
d['tue'].append(8)
print(res)
```

What is the output of the above script?

- (a) The script will generate an error.
- (b) dict_values([[1, 2, 3, 4], [2, 4, 6]])
- (c) None of the other answers is correct.
- (d) dict_values([[1, 2, 3], [2, 4, 6]])
- (e) dict_values([[1, 2, 3], [2, 4, 6, 8]])
- (f) dict_values([[1, 2, 3, 4], [2, 4, 6, 8]])

Domanda 10

Risposta non data

Punteggio max.: 1,50

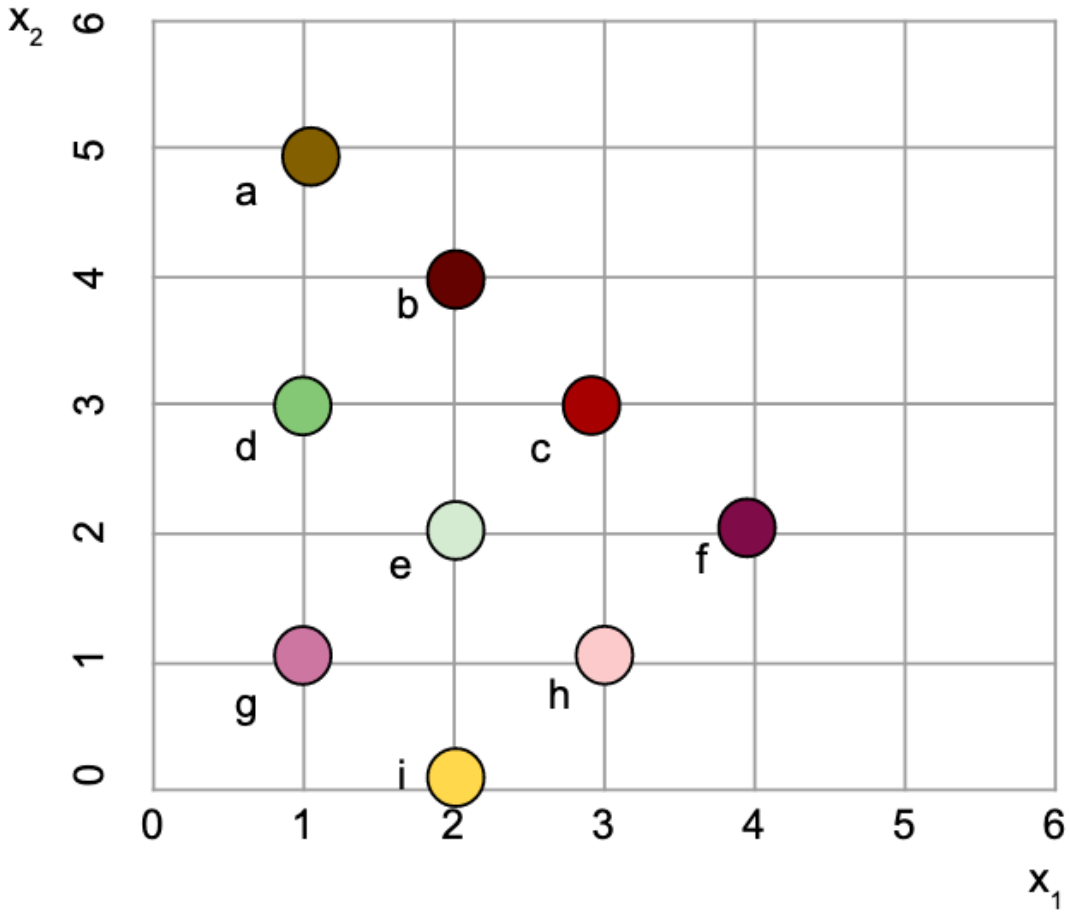
1.5 points - -15% penalty for a wrong answer

The Mean Squared Error between two vectors $z, z' \in \mathbb{R}^n$ is defined as:

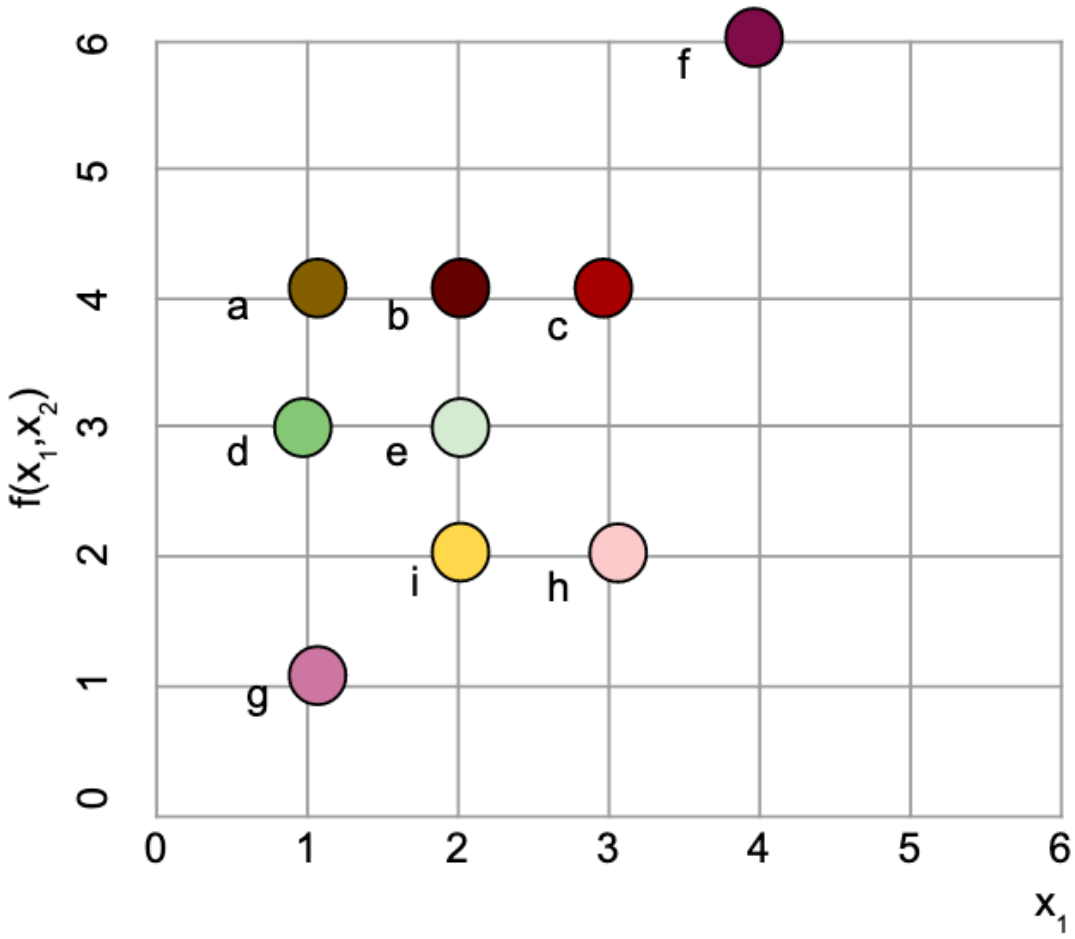
$$MSE(z, z') = \frac{1}{n} \sum_i (z_i - z'_i)^2$$

Consider the points in the figures below with features x_1, x_2 and ground truth values $f(x_1, x_2)$.

Input space



Ground truth f(x_1, x_2)



Apply the following regression model to all the data points:

$$f'(x_1, x_2) = 0.5x_1 + 0.5x_2 + 1$$

Write in the answer box the value of the MSE between the prediction $f'(x_1, x_2)$ and the ground truth $f(x_1, x_2)$

Use the following syntax:
mse=value

Example:
mse=3

Domanda 11

Risposta non data

Non valutata

This is not a question

You can use the text area below to write any note or draft (e.g. intermediate steps of an exercise).

Any text written below will not be considered toward the correction of the exam.

Domanda 12

Risposta non data

Non valutata

Withdrawal from the exam

THIS IS NOT AN EXAM QUESTION.

If you wish to withdraw from the exam, please select the option "Withdraw". The exam will not be corrected.

Otherwise, you may leave this answer blank (or select "DO NOT withdraw").

- (a) Withdraw
- (b) **DO NOT** withdraw