Domanda 1

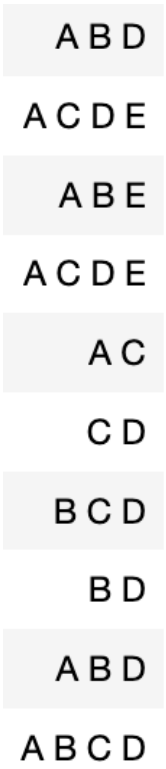
Risposta non data

Punteggio max.: 3,00

3 points (no penalty for a wrong answer)

```
An itemset X is closed if none of its supersets has the same support as X.
```

Given the transactional dataset shown in the figure below, apply the Apriori algorithm to extract all frequent itemsets.



The value of minsup is 2 (an itemset is frequent if it appears in at least 2 transactions).

An itemset is considered to be frequent if its support count is equal to or higher than the minsup.

1. List all frequent itemsets having length 2, **along with their support count**.
2. List all candidate itemsets of length 3 that have been generated by Apriori **after the join and prune steps**, before counting their support in the database.
3. List all frequent itemsets that are **not** closed, along with their support count.

Use the following notation:

```
A1) { ... list of itemsets w/ support count ... }
A2) { ... list of itemsets ... }
A3) { ... list of itemsets w/ support count ... }
```

For example

```
A1) { ab: 2, ac: 3, ad: 2 }
A2) { abc, abd, abe }
A3) { ab: 2, ad: 2, bce: 3 }
```

Domanda 2

Risposta non data

Punteggio max.: 1,00

1 point (-15% penalty for a wrong answer)

A random forest (RF) and a decision tree (DT) classifiers have been trained on a dataset D.

You accidentally train the random forest using only 1 decision tree.

Which of the following statements is correct? Choose all correct answers (multiple answers may be correct)

Note: when referring to "RF" in the answers, it means the specific 1-tree random forest mentioned above.

Scegli una o più alternative:

- a. None of the other statements is correct
- b. The RF and the DT will produce different results, since the voting scheme of the RF will affect its predictions
- c. The RF and the DT will produce the same results
- d. The RF and the DT will produce different results, since the pool of features used changes for each split of DT
- e. The RF and the DT will produce different results, since the data used for training them will differ
- f. The RF and the DT will produce different results, since the pool of features used changes for each split of RF

Domanda 3

Risposta non data

Punteggio max.: 1,50

1.5 points (-15% penalty for a wrong answer)

What is the output of the following code? (you can assume infinite precision in the representation of floating point numbers)

```
x = np.random.random((128, 16))
mu = x.mean(axis=0)
sigma = x.std(axis=0)

x2 = (x - mu) / sigma

w = np.random.random(16)
b = np.random.random()

y = (x2 * w).sum(axis=1) + b

print(x2.mean(axis=0).mean(), y.shape)
```

Scegli un'alternativa:

- a. 0 (16,)
- b. 0.5 (16,)
- c. An error occurs
- d. 1 (128, 16)
- e. 0 (128, 16)
- f. 1 (128,)
- g. None of the other answers is correct
- h. 1 (16,)
- i. 0 (128,)
- j. 0.5 (128,)
- k. 0.5 (128, 16)

Domanda 4

Risposta non data

Punteggio max.: 1,50

1.5 points (no penalty for a wrong answer)

For a vector of predictions $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ and ground truths $y = (y_1, y_2, \ldots, y_n)$, the R2 score is defined as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Where $\bar{y} = \frac{1}{n} \sum_i y_i$.

A training set is comprised of 10,000 15-dimensional points and 1 target variable. The target variable is continuous and has mean = 5.9 and standard deviation 3.4. The training set is used to train a regression algorithm to predict the target variable.

A separate test set comprised of 5,000 points is next used for assessing the quality of the regression model. The target variable in the test set can be assumed to have the same distribution as the target variable in the training set.

The sum of squared errors for the model is of 14,807.7651.

Based on the available information, what is the R2 of the model, as evaluated on the test set?

Write the answer in the box below. If not enough information is provided, write "not enough information available".

Note on numerical representations

You can report the results as either fractions or numbers.

- When reporting fractions, use the syntax N/D, where both N and D are numerical values.
- When reporting numbers, use at least 4 significant figures (unless fewer figures are required to represent a number with full precision).

`

Valid representations:

```
1/2 or 0.5
1/7 or 0.1429 or 0.142857
```

Invalid representations:

```
(2+4+9)/(3*10) (to represent 1/2)
0.14 or 0.143 (to represent 1/7)
```

Domanda 5

Risposta non data

Punteggio max.: 1,50

1.5 points (no penalty for a wrong answer)

The following are the pandas DataFrames *df1* and *df2,* respectively (note the column names):

	height	weight	width
0	2	1	2
1	5	7	2
2	6	3	5
3	1	8	7
4	8	1	6

	width	height	weight
0	8	5	3
1	1	3	6
2	9	4	6
3	2	3	1
4	9	9	9

What are the contents of *df5*, after executing the following snippet of code?

```
df3 = (df1 + df2)
df4 = df3 - df3.min()
df5 = df4.loc[:3]
```

Write your answer in the box below, specifying:

- Type (DataFrame or Series)
- Columns (for DataFrames only)
- Index
- Data (the contents of the Series/DataFrame)

If the execution of the code results in an error, write “an error occurs” instead.

For "Data", you can separate multiple columns (in the case of DataFrames) using a space. You can separate rows using newlines.

In the case of a multi-level index/column, you must represent them in the same way as with the data (separating with spaces and newlines). You need to repeat the outermost index/column whenever needed.

The following are some examples of the representation you should use.

Example 1

DataFrame

	a	b	c
0	3	3	2
1	2	4	4

Expected representation

```
Type
DataFrame

Columns
a b c

Index
0 1

Data
3 3 2
2 4 4
```

Example 2

Multi-index/column DataFrame

		W		X
		a	b	c
y	**0**	2	3	3
	1	4	3	4
z	**0**	3	1	4
	1	3	3	4

Expected representation

```
Type
DataFrame

Columns
W W X
a b c

Index
y y z z
0 1 0 1

Data
2 3 3
4 3 4
3 1 4
3 3 4
```

Example 3

Series

```
a    11
b     7
c     5
d    -1
```

Expected representation

```
Type
Series

Index
a b c d

Data
11 7 5 −1
```

Domanda 7

Risposta non data

Punteggio max.: 2,00

2 points (no penalty for a wrong answer)

You are given the following distance matrix among 5 points, a, b, c, d, e.

	a	b	c	d	e
a	0	13	22	17	23
b	13	0	16	7	21
c	22	16	0	24	6
d	17	7	24	0	20
e	23	21	6	20	0

Apply agglomerative hierarchical clustering using complete (MAX) linkage.

Write in the box below the clusters obtained at each step of the clustering.

Use a separate line to represent the state after each step. Separate the clusters obtained using a vertical bar (|).

Start from the case where each point belong to its own cluster. End with the case where all points belong to a single cluster.

For example,

```
a | b | c | d | e
a b | c | d | e
a b c | d | e
a b c d | e
a b c d e
```

In this example, the steps of the clustering first merge a with b, then (a, b) with c, then (a, b, c) with d, and finally (a, b, c, d) with e.

Domanda 8

Risposta non data

Punteggio max.: 1,50

1.5 points (-15% penalty for a wrong answer)

You are given a dataset X with target classes y. Both X and y are represented as NumPy arrays. The contents of X and y are unknown, you only know the following:

- X.ndim == 2
- y.ndim == 1
- X.shape[0] == y.shape[0]
- X.shape[0] > 5
- X and y do not contain any missing values
- X contains real numbers in the range [0, 100]
- y contains one of 10 classes encoded as integers from 0 to 9

The following piece of code is used to find the optimal set of hyperparameters for a decision tree model.

```
from sklearn.model_selection import KFold, ParameterGrid
from sklearn.tree import DecisionTreeClassifier

params = {
    "max_depth": [5, 10, None],
    "criterion": ["gini", "entropy", "log_loss"],
    "min_samples_split": [2, 5]
}

for config in ParameterGrid(params):
    for train_ndx, test_ndx in KFold(5).split(X, y):
        clf = DecisionTreeClassifier(max_depth=config["max_depth"], criterion=config["criterion"],
min_samples_split=config["min_samples_split"])
        clf.fit(X[train_ndx], y[train_ndx])
        y_pred = clf.predict(X[test_ndx])
        # ... code for handling the evaluation
```

How many decision trees are trained based on the above code?

Scegli un'alternativa:

- a. 180
- b. 1
- c. 5
- d. 90
- e. None of the other answers is correct
- f. An error occurs
- g. 18
- h. Not enough information is provided to address the question
- i. 40

Domanda 9

Risposta non data

Punteggio max.: 1,50

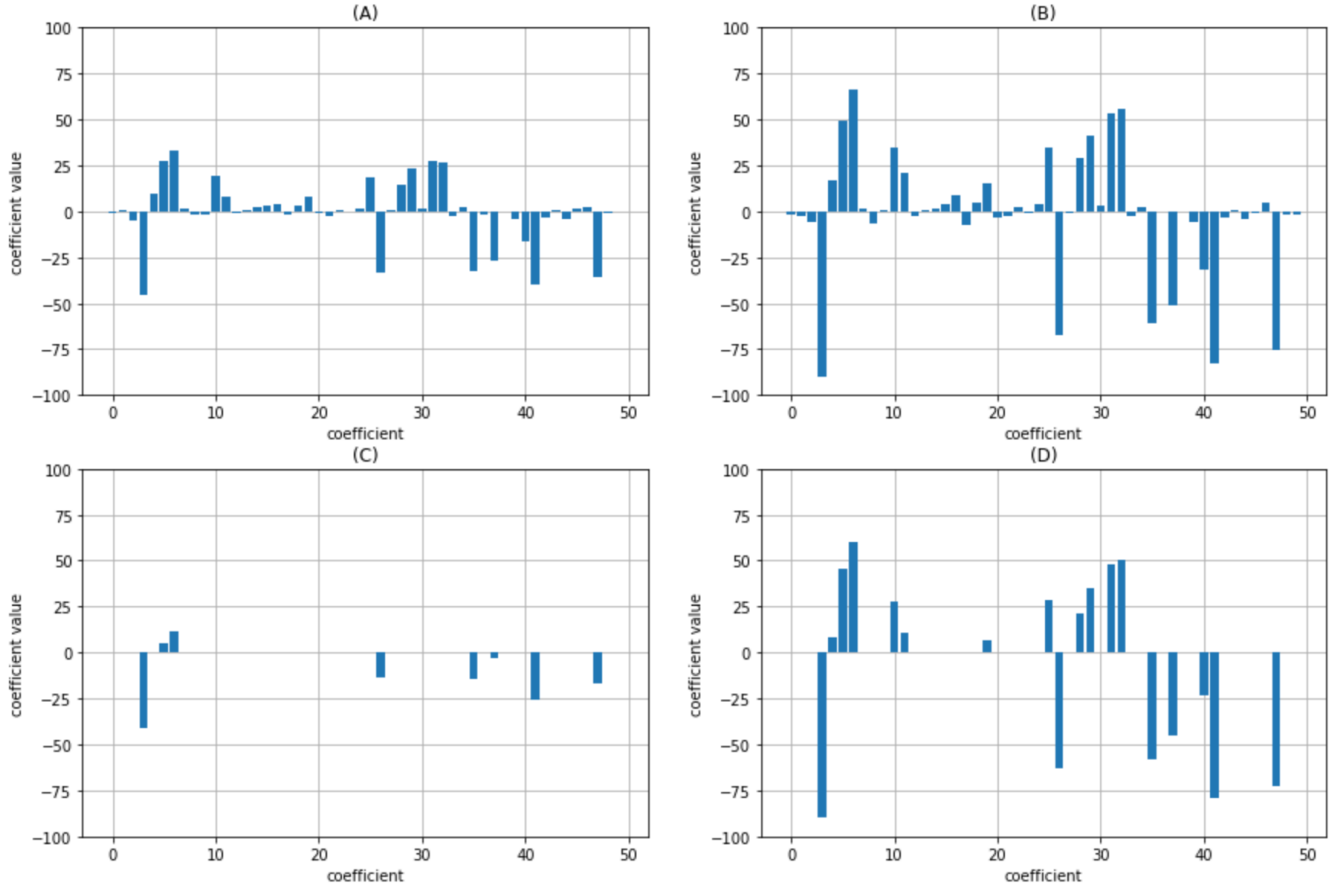
1.5 points (-15% penalty for a wrong answer)

The weights w of a Ridge model are chosen so as to minimize $MSE + \lambda||w||_2^2$
The weights w of a Lasso model are chosen so as to minimize $MSE + \lambda||w||_1$

Two Ridge models and two Lasso models have been trained on the same dataset comprised of 1,000 50-dimensional points. In particular, the following are the characteristics of the four models:

- Ridge, $\lambda = 10$
- Ridge, $\lambda = 100$
- Lasso, $\lambda = 1$
- Lasso, $\lambda = 5$

After training, you plot the weights (coefficients) learned by the four models as bar charts. The following are the plots you obtain.



You labelled the four plots with letters (A, B, C, D respectively). Which model corresponds to which plot?

Scegli un'alternativa:

a. (A) Ridge, $\lambda = 10$
(B) Ridge, $\lambda = 100$
(D) Lasso, $\lambda = 1$
(C) Lasso, $\lambda = 5$

b. (A) Ridge, $\lambda = 10$
(B) Ridge, $\lambda = 100$
(C) Lasso, $\lambda = 1$
(D) Lasso, $\lambda = 5$

c. (A) Ridge, $\lambda = 10$
(C) Ridge, $\lambda = 100$
(B) Lasso, $\lambda = 1$
(D) Lasso, $\lambda = 5$

d. (C) Ridge, $\lambda = 10$
(B) Ridge, $\lambda = 100$
(A) Lasso, $\lambda = 1$
(D) Lasso, $\lambda = 5$

e. (B) Ridge, $\lambda = 10$
(A) Ridge, $\lambda = 100$
(C) Lasso, $\lambda = 1$
(D) Lasso, $\lambda = 5$

f. (A) Ridge, $\lambda = 10$
(D) Ridge, $\lambda = 100$
(B) Lasso, $\lambda = 1$
(C) Lasso, $\lambda = 5$

g. (D) Ridge, $\lambda = 10$
(B) Ridge, $\lambda = 100$
(C) Lasso, $\lambda = 1$
(A) Lasso, $\lambda = 5$

h. (B) Ridge, $\lambda = 10$
(C) Ridge, $\lambda = 100$
(D) Lasso, $\lambda = 1$
(A) Lasso, $\lambda = 5$

i. None of the other answers is correct.

j. (B) Ridge, $\lambda = 10$
(C) Ridge, $\lambda = 100$
(A) Lasso, $\lambda = 1$
(D) Lasso, $\lambda = 5$

Domanda 10

Risposta non data

Punteggio max.: 1,00

1 point (-15% penalty for a wrong answer)

```
An itemset is maximal if it is frequent and none of its supersets is frequent
```

After extracting the frequent itemsets from a transactional dataset, you find that -- among others -- the itemsets ABC and BD are maximal.

Which of the following statements is correct?

Scegli un'alternativa:

- a. None of the other statements is correct
- b. ABCD is maximal
- c. ABCD is frequent
- d. ABCD is frequent only if BCD is frequent
- e. AC is frequent
- f. ABD is frequent
- g. AC is maximal
- h. ABD is maximal

Domanda 11

Risposta non data

Punteggio max.: 2,00

2 points (no penalty for a wrong answer)

Given a set of documents $D = \{d_1, d_2, \dots\}$ and a set of terms $T = \{t_1, t_2, \dots\}$, the tf-idf for a term $t \in T$ in a document $d \in D$ is defined as:

$tf - idf(t, d, D) = tf(t, d) \cdot idf(t, D)$

A possible definition for $tf(t, d)$ is the number of occurrences of t within d.

A possible definition of $idf(t, D)$ is the following:

$idf(t, D) = log \frac{|D|}{|\{d \in D : t \in d\}|}$

You are given the following collection of documents:

```
<term 0> <term 0> <term 0> <term 3> <term 0>
<term 0> <term 1> <term 0> <term 0>
<term 3> <term 3> <term 2> <term 2>
```

Where each line represents a document and each term is represented in the form <term X>.

Compute the tf-idf matrix for all terms/documents.

Report the result in the box below, listing one document for each line and one term for each column.

For the documents, use the same order in which the documents appear in the above list. For the terms, use their natural order (<term 0>, <term 1>, etc).

```
tfidf_term_0_doc_0 tfidf_term_1_doc_0 tfidf_term_2_doc_0 tfidf_term_3_doc_0
tfidf_term_0_doc_1 tfidf_term_1_doc_1 tfidf_term_2_doc_1 tfidf_term_3_doc_1
tfidf_term_0_doc_2 tfidf_term_1_doc_2 tfidf_term_2_doc_2 tfidf_term_3_doc_2
```

For example,

```
0.1234 0.4321 0.5555 0.6667
0.6667 0.5555 0.4321 0.1234
0.1212 0.2121 0.3333 0.3333
```

Use the natural logarithm (ln) for the computation of the idf.

Note on numerical representations

You can report the results as either fractions or numbers.

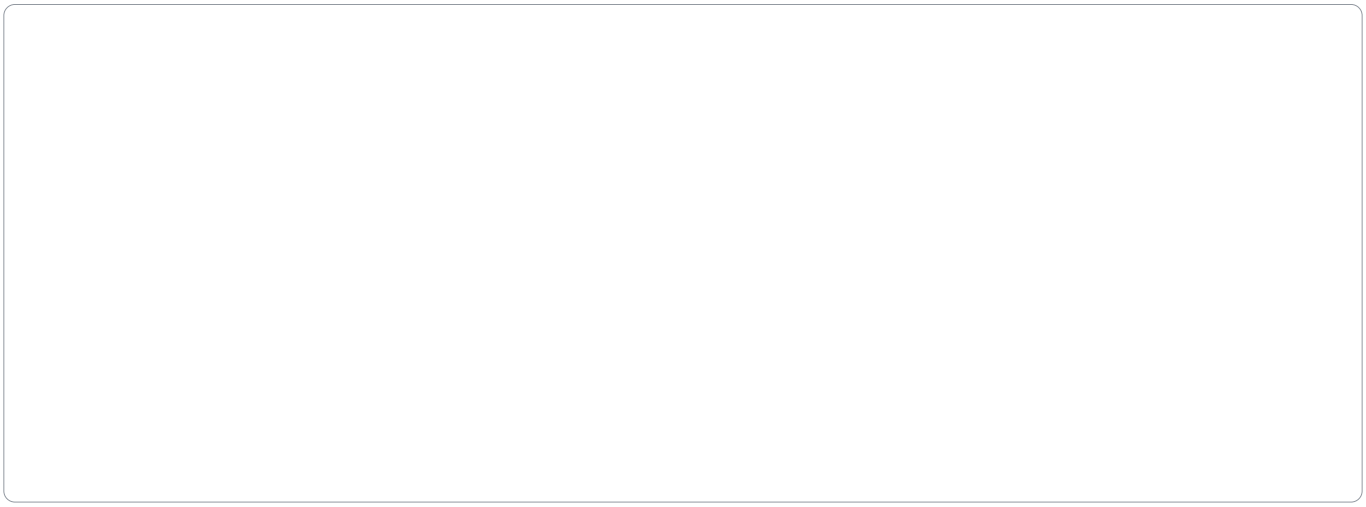
- When reporting fractions, use the syntax N/D, where both N and D are numerical values.
- When reporting numbers, use at least 4 significant figures (unless fewer figures are required to represent a number with full precision).
- For this exercise you are required to represent results that contain logarithms. You can report the results in the form A * log (B). Both A and B must follow the above rules

Valid representations:

```
1/2 or 0.5
1/7 or 0.1429 or 0.142857
```

Invalid representations:

```
(2+4+9)/(3*10) (to represent 1/2)
0.14 or 0.143 (to represent 1/7)
```



Domanda 12

Risposta non data

Punteggio max.: 2,50

2.5 points (no penalty for a wrong answer)

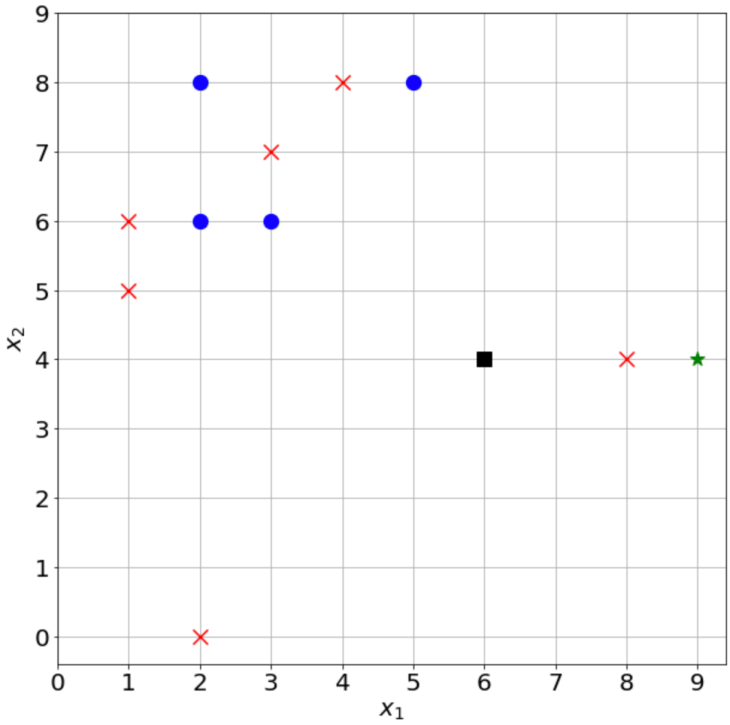
The KNN algorithm can assign a class probability based on the votes assigned to each class. If $vote_Y(x)$ represents the unnormalized vote assigned to class Y for sample x, then, for any class $c \in C$:

$p(c|x) = \frac{vote_c(x)}{\sum_{i \in C} vote_i(x)}$

The L_∞ (or Chebyshev) distance between points $x = (x_1, x_2, \ldots, x_d)$ and $y = (y_1, y_2, \ldots, y_d)$ in d dimensions is defined as:

$L_\infty(x, y) = max_i|x_i - y_i|$

The figure below shows a 2-dimensional dataset. Points in this dataset belong to either of two classes: circles (in blue) or crosses (in red).



You are given two test points to be labelled, A (green star) and B (black square). The K-NN algorithm (with L_∞ distance and K = 3) is used to label A and B.

Answer the following questions.

Question 1) What is the label assigned to A and B, when using uniform votes?
Question 2) What are the class probabilities for A and B, if each vote is weighted with a weight $w = \frac{1}{1+distance}$?

Use the following notation:

```
Answer 1) A=class_for_A, B=class_for_B
Answer 2)
A: cross=p(cross|A), circle=p(circle|A)
B: cross=p(cross|B), circle=p(circle|B)
```

For example,

```
Answer 1) A=circle, B=circle
Answer 2)
A: cross=1/2, circle=0.5
B: cross=0.9091, circle=0.0909
```

Note on numerical representations

You can report the results as either fractions or numbers.

- When reporting fractions, use the syntax N/D, where both N and D are numerical values.
- When reporting numbers, use at least 4 significant figures (unless fewer figures are required to represent a number with full precision).

`

Valid representations:

```
1/2 or 0.5
1/7 or 0.1429 or 0.142857
```

Invalid representations:

```
(2+4+9)/(3*10) (to represent 1/2)
0.14 or 0.143 (to represent 1/7)
```

Domanda 13

Risposta non data

Punteggio max.: 1,00

1 point (-15% penalty for a wrong answer)

For the DBSCAN algorithm, the *epsilon* parameter defines the radius considered when searching for neighboring points, whereas *min_points* represents the number of neighboring points expected to be found within a radius of *epsilon* for a point to be considered *core*.

Which of the following statements on DBSCAN is correct?

Scegli un'alternativa:

- a. For a fixed value of min_points, increasing the value of epsilon generally increases the number of core points found.
- b. For a fixed value of epsilon, increasing min_points generally increases the number of core points found
- c. The values of min_points and epsilon do not affect the total number of clusters found
- d. Only epsilon affects the total number of clusters found
- e. For a fixed value of min_points, increasing the value of epsilon generally increases the number of border points found.
- f. None of the other statements is correct
- g. Only min_points affects the total number of clusters found