ML - EXAMS

Rules

ANSWER SHEET

- **Make sure you read the instructions carefully.**
- Total time: 50 minutes.
- Usage of texts, notes, or digital devices is forbidden.
- Questions can have multiple correct answers.
- Wrong answers are attributed a negative score (half the score of a correct answer).
- Completely blacken the box beside a correct answer **in the answer sheet ONLY**. Answers indicated in other sheets will be ignored. Use a white-out marker to correct and unmark previously marked boxes.

  DOs: ☐ ■   **DON'Ts:** ☒ ☒ ☑   **Absolute DON'Ts:** ■ NO

- No questions are allowed during the exam. Comments can be included in the last box. If your comment refers to a question, make sure you reference the question (for example, "In question 1, answer B, ...").
- It is your responsibility to write in a clear and understandable way.
- **DO NOT** black out the boxes in the top right corner of open text questions.
- **To withdraw** from the exam, completely blacken out the box associated to Question 11 in the answer sheet.

Rules

- Please encode your **codice persona** (8 digits) below, blacking out one digit per column. For instance, if your codice persona is 12345678, the "1" digit must be blacked out in the leftmost column, and the "8" digit must be blacked out in the rightmost column.

0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9

- Write your first and last names in the box below.

Last name: ....

First name: ....

Rules

Answers:

QUESTION 1: A B C D E F

QUESTION 2: A B C D E F

QUESTION 3: A B C D E F

QUESTION 4: A B C D E F

QUESTION 5: A B C D E F

QUESTION 6: A B C D E F

QUESTION 7: A B C D E F

QUESTION 8: A B C D E F

QUESTION 9: A B C D E F

QUESTION 10: C

QUESTION 11: A

Question 1 ♣ Which one(s) among the following answers is(are) correct?

A	Filter feature selection methods are model-dependent.
B	Data discretization is a data reduction technique.
C	Mean absolute deviation is a measure of dispersion.
D	Multivariate exploratory analysis has the downside of not exploiting visualization techniques.
E	Min-max standardization is robust to outliers.
F	*None of these answers are correct.*

Question 1 ♣ Which one(s) among the following answers is(are) correct?

A Filter feature selection methods are model-dependent.

■ Data discretization is a data reduction technique.

■ Mean absolute deviation is a measure of dispersion.

D Multivariate exploratory analysis has the downside of not exploiting visualization techniques.

E Min-max standardization is robust to outliers.

F *None of these answers are correct.*

Question 1 ♣ Which one(s) among the following answers is(are) correct?

A	Filter feature selection methods are model-dependent.
■	Data discretization is a data reduction technique.
■	Mean absolute deviation is a measure of dispersion.
D	Multivariate exploratory analysis has the downside of not exploiting visualization techniques.
E	Min-max standardization is robust to outliers.
F	*None of these answers are correct.*

B. Data discretization transforms continuous variables into discrete intervals or bins. This process simplifies the data representation, effectively reducing the amount of data while preserving its essential structure for analysis.

C. Mean absolute deviation (MAD) quantifies the average absolute deviation of data points from their mean, representing data variability or dispersion. It's a robust measure compared to variance when the dataset contains outliers.

Question 2 ♣ Referring to classification, which one(s) among the following answers is(are) correct?

A   In the F-measure, increasing β increases the relative importance of the recall.

B   Classification trees can perform Multi-class classification by resorting to techniques such as one-vs-all or round-robin.

C   Referring to ROC curve, the trajectory of the curve is based on the classifier scoring.

D   The Gini index takes value one for a maximum heterogeneity.

E   Naive Bayes classifiers are based on hypotheses of the probability $P(\mathbf{x}|y)$.

F   *None of these answers are correct.*

Question 2 ♣ Referring to classification, which one(s) among the following answers is(are) correct?

A In the F-measure, increasing β increases the relative importance of the recall.

B Classification trees can perform Multi-class classification by resorting to techniques such as one-vs-all or round-robin.

■ Referring to ROC curve, the trajectory of the curve is based on the classifier scoring.

D The Gini index takes value one for a maximum heterogeneity.

■ Naive Bayes classifiers are based on hypotheses of the probability $P(\mathbf{x}|y)$.

F *None of these answers are correct.*

Question 2 ♣ Referring to classification, which one(s) among the following answers is(are) correct?

A In the F-measure, increasing β increases the relative importance of the recall.

B Classification trees can perform Multi-class classification by resorting to techniques such as one-vs-all or round-robin.

■ Referring to ROC curve, the trajectory of the curve is based on the classifier scoring.

D The Gini index takes value one for a maximum heterogeneity.

■ Naive Bayes classifiers are based on hypotheses of the probability $P(\mathbf{x}|y)$.

F *None of these answers are correct.*

C. The Receiver Operating Characteristic (ROC) curve depends on the classifier's scoring method, such as probabilities or decision values, to determine the trade-off between true positive and false positive rates at different thresholds.

E. Naive Bayes classifiers rely on the assumption of feature independence, estimating the probability of each feature given the class (P(x|y)P(x|y)P(x|y)) to compute the overall likelihood using Bayes' theorem.

Question 3 ♣ Referring to regression, which one(s) among the following answers is(are) correct?

A Extracting the feature "BMI" ($\frac{kg}{m^2}$) from features "height" (m) and "weight" (kg) never helps a linear regression model.

B The QQ plot can be used to verify if the residuals have zero mean.

C In Bagging methods, the final prediction is the one provided by the last generated (weak) learner.

D Support Vector Regression cannot use kernels other than the linear one.

E A model having a value of the MAE on the training set considerably lower than MAE on the test set is likely to be underfitted.

F *None of these answers are correct.*

Question 3 ♣ Referring to regression, which one(s) among the following answers is(are) correct?

A Extracting the feature "BMI" ($\frac{kg}{m^2}$) from features "height" (m) and "weight" (kg) never helps a linear regression model.

B The QQ plot can be used to verify if the residuals have zero mean.

C In Bagging methods, the final prediction is the one provided by the last generated (weak) learner.

D Support Vector Regression cannot use kernels other than the linear one.

E A model having a value of the MAE on the training set considerably lower than MAE on the test set is likely to be underfitted.

■ *None of these answers are correct.*

Question 4 ♣ Referring to the clustering problem, which one(s) among the following answers is(are) correct?

A Contrary to K-means, the output of K-medoids does not depend on the starting condition.

B Complete clustering techniques cannot be used for outlier detection.

C Mahalanobis distance is a generalization of the Euclidean distance.

D Fuzzy clustering methods assign the observations to the cluster with a weight between 0 and 1.

E Regarding hierarchical methods, since the complexity of computing cluster dissimilarity grows with the number of observations when the number of observations is large, it will be convenient to generate all possible clusters and find the one having maximum quality.

F *None of these answers are correct.*

Question 4 ♣ Referring to the clustering problem, which one(s) among the following answers is(are) correct?

A Contrary to K-means, the output of K-medoids does not depend on the starting condition.

B Complete clustering techniques cannot be used for outlier detection.

■ Mahalanobis distance is a generalization of the Euclidean distance.

■ Fuzzy clustering methods assign the observations to the cluster with a weight between 0 and 1.

E Regarding hierarchical methods, since the complexity of computing cluster dissimilarity grows with the number of observations when the number of observations is large, it will be convenient to generate all possible clusters and find the one having maximum quality.

F *None of these answers are correct.*

C. Mahalanobis distance accounts for correlations between variables, unlike Euclidean distance, making it more robust in multidimensional clustering.

D. Fuzzy clustering assigns a membership degree (0 to 1) to each observation for all clusters, reflecting partial belonging

Question 5 ♣ Referring to association rules, which one(s) among the following answers is(are) correct?

A The confidence of the rule $\{A \Rightarrow B\}$ expresses the empirical conditional probability of finding the itemset B in a transaction given that items in A are present.

B The rules $\{A \Rightarrow B\}$ and $\{B \Rightarrow A\}$ have the same value of support and lift.

C If a rule has maximum confidence then its body is present in all the transactions.

D Let C be an itemset containing A, then if the rule $\{A \Rightarrow B\}$ is strong then the rule $\{C \Rightarrow B\}$ will be also strong.

E The support of the rule $\{A \Rightarrow B\}$ expresses the frequency of B conditional to A.

F *None of these answers are correct.*

Question 5 ♣ Referring to association rules, which one(s) among the following answers is(are) correct?

■ The confidence of the rule $\{A \Rightarrow B\}$ expresses the empirical conditional probability of finding the itemset B in a transaction given that items in A are present.

■ The rules $\{A \Rightarrow B\}$ and $\{B \Rightarrow A\}$ have the same value of support and lift.

C If a rule has maximum confidence then its body is present in all the transactions.

D Let C be an itemset containing A, then if the rule $\{A \Rightarrow B\}$ is strong then the rule $\{C \Rightarrow B\}$ will be also strong.

E The support of the rule $\{A \Rightarrow B\}$ expresses the frequency of B conditional to A.

F *None of these answers are correct.*

A. Confidence measures how often B appears in transactions containing A, which is equivalent to conditional probability P(B|A).

B. Support and lift depend on the co-occurrence of A and B, making them symmetric regardless of rule direction.

Question 6   The figure depicts the code used in a regression task applied to a dataset.

- Identify the regression model used in this example.

- Describe each one of the parameters used in both the model parameter set and the grid search definition, explaining their role clearly.

- There is a clear error in the code. Identify the error and motivate your answer.

```python
from sklearn.ensemble import
from sklearn.model_selection import GridSearchCV
import numpy as np

regressor =
parameters = {"n_estimators":[5,10,100], "criterion": ['mae'],
              "max_depth":[2,10], "min_samples_leaf": [0.1,0.3]}

gs = GridSearchCV(regressor, parameters, scoring='f1', cv=3) |
```

The code uses an ensemble model **Random Forest Regressor** or a similar estimator

Question 6   The figure depicts the code used in a regression task applied to a dataset.

n_estimators: Number of estimators (e.g., decision trees) in the ensemble. Higher values typically improve performance but increase computational cost. Values being tested: [5, 10, 100].

criterion: The function used to measure the quality of a split. The value 'mae' stands for Mean Absolute Error, commonly used in regression tasks.

max_depth: The maximum depth of each tree in the ensemble. A smaller value prevents overfitting, while a larger depth allows for more complex patterns. Values tested: [2, 10].

min_samples_leaf: The minimum number of samples required to be at a leaf node. Smaller values lead to more splits, potentially increasing overfitting. Values tested: [0.1, 0.3] (likely as fractions of the dataset size).

scoring: Specifies the metric used to evaluate performance during cross-validation. 'f1' is generally for classification tasks and is incompatible with regression problems.

cv: The number of cross-validation splits. In this case, it's set to 3.
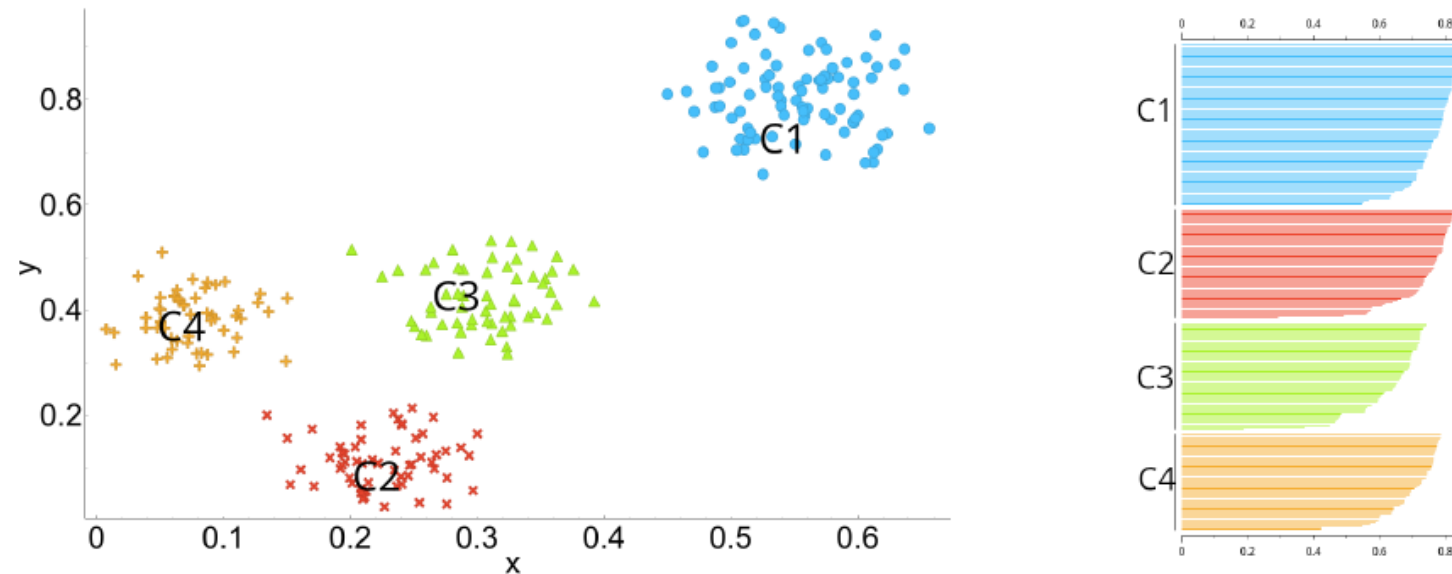
Incompatibility Between Scoring and Problem Type:
   The scoring metric 'f1' is specific to classification tasks. For regression tasks, metrics like 'neg_mean_absolute_error', 'neg_mean_squared_error', or 'r2' should be used.
   Using 'f1' would likely result in an error or invalid evaluation results because it does not apply to continuous target variables.

Question 7 ♣ The figures depict the results obtained after applying a clustering algorithms on a dataset containing two variables. The partition of the observations is depicted in the left figure: the symbol of each observation indicates the cluster to which it belongs. On the other hand, the right figure represents the silhouette scores for the resulting clusters.
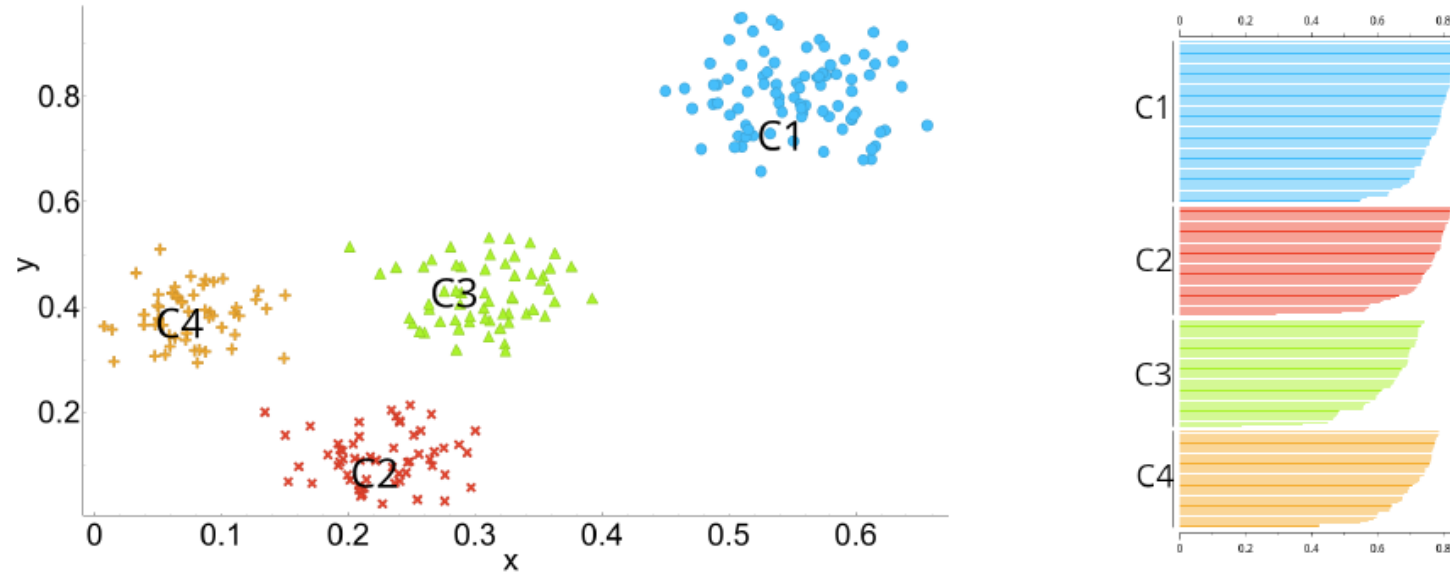
Regarding the figures, which one(s) among the following answers is(are) correct?

| A | If we remove the observations belonging to the third cluster (C3) then the value of the silhouette of the remaining observations will remain the same. |

| B | The clustering method applied to the dataset could be hierarchical agglomerative with a complete(max) linkage. |

| C | The clustering method applied to the dataset could be k-means. |

| D | To obtain a better mean silhouette value, one could increase the number of clusters. |

| E | If we remove the observations belonging to the first cluster (C1) then the value of the silhouette of the remaining observations will remain the same. |

| F | *None of these answers are correct.* |

Question 7 ♣ The figures depict the results obtained after applying a clustering algorithms on a dataset containing two variables. The partition of the observations is depicted in the left figure: the symbol of each observation indicates the cluster to which it belongs. On the other hand, the right figure represents the silhouette scores for the resulting clusters.

Regarding the figures, which one(s) among the following answers is(are) correct?

A   If we remove the observations belonging to the third cluster (C3) then the value of the silhouette of the remaining observations will remain the same.

■   The clustering method applied to the dataset could be hierarchical agglomerative with a complete(max) linkage.

■   The clustering method applied to the dataset could be k-means.

D   To obtain a better mean silhouette value, one could increase the number of clusters.

■   If we remove the observations belonging to the first cluster (C1) then the value of the silhouette of the remaining observations will remain the same.

F   *None of these answers are correct.*

A	If we remove the observations belonging to the third cluster (C3) then the value of the silhouette of the remaining observations will remain the same.
■	The clustering method applied to the dataset could be hierarchical agglomerative with a complete(max) linkage.
■	The clustering method applied to the dataset could be k-means.
D	To obtain a better mean silhouette value, one could increase the number of clusters.
■	If we remove the observations belonging to the first cluster (C1) then the value of the silhouette of the remaining observations will remain the same.
F	*None of these answers are correct.*

B: Hierarchical clustering with complete linkage minimizes the maximum distance between points in different clusters, aligning with the given silhouette pattern.

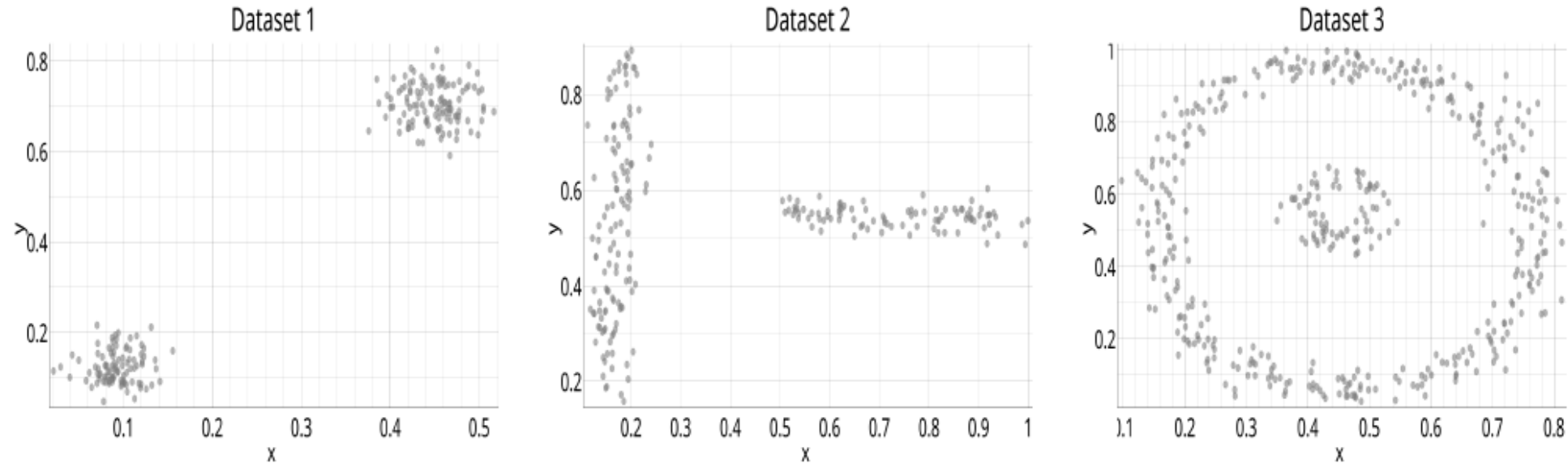C: K-means produces distinct partitions and silhouette scores, consistent with the example.

E: Cluster C1 represents a well-defined group with high internal cohesion and clear separation from other clusters, removing it could reduce the overall quality of the clustering evaluation metrics, such as **silhouette scores**, because the dataset would lose a distinct and well-formed cluster.

Question 8 ♣ The following figures depict three different datasets, each one containing two features. In each figure, points represent the observation in the dataset, and their positions correspond to the values of the two variables. If in each of the datasets we apply Principal Component Analysis(PCA), which one(s) among the following answers is(are) correct?
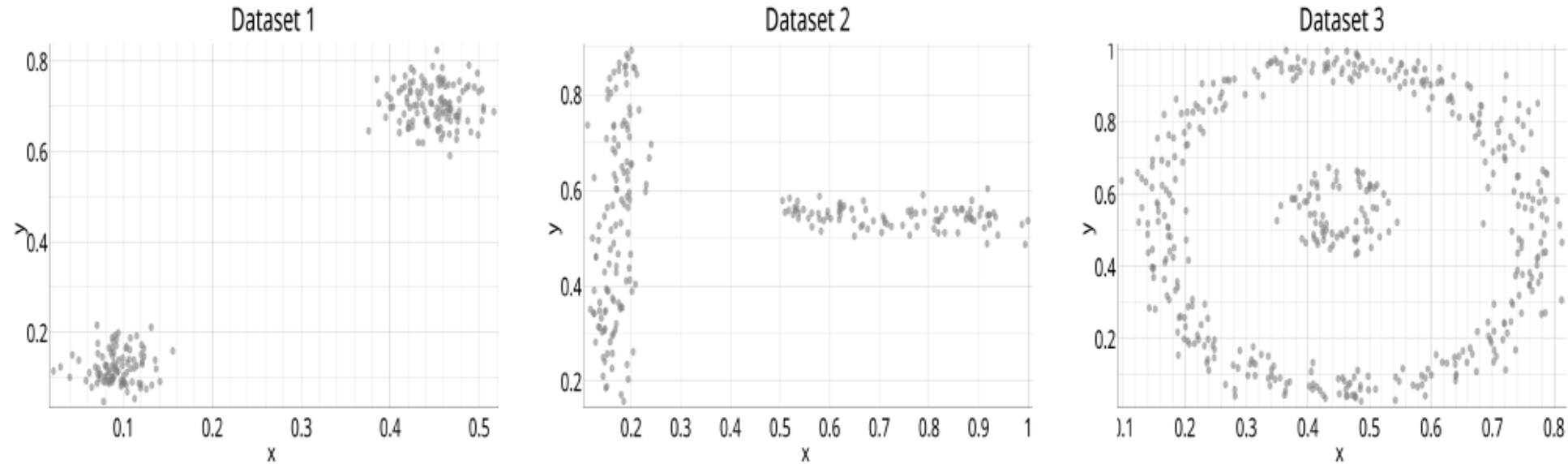
A In the third dataset the principal components could not be orthogonal.

B In the second dataset, the first principal component explains a large percentage of the total variance.

C In the second dataset, the principal components' directions are similar to the main direction of the system of coordinates.

D In the first dataset, the first principal component explains a large percentage of the total variance.

E In the third dataset, the first principal component explains a large percentage of the total variance.
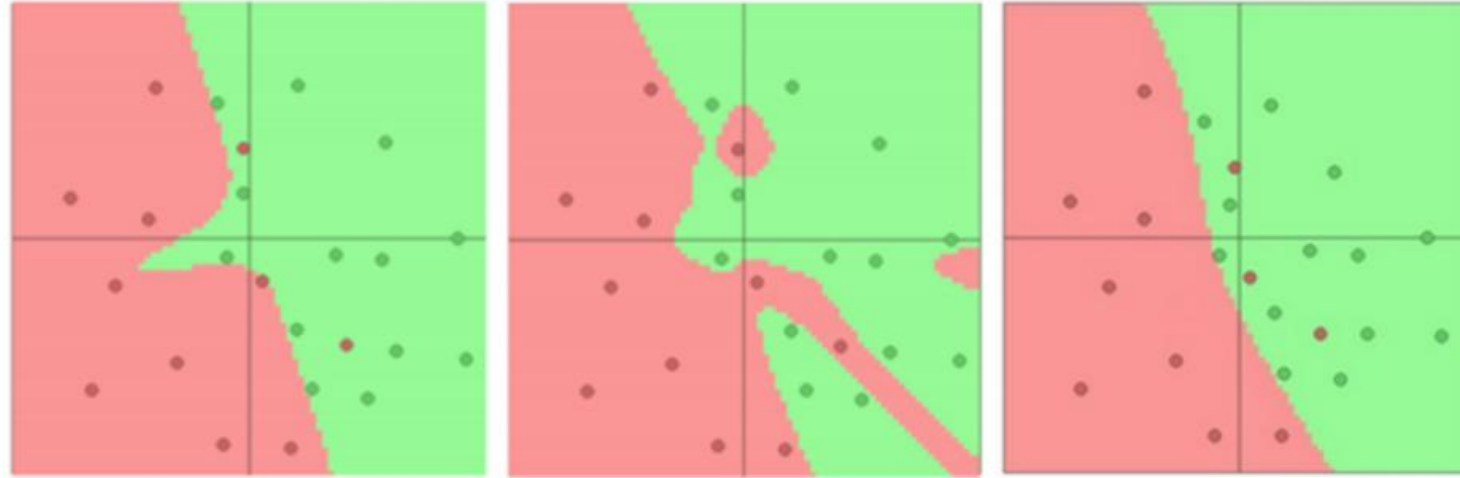
F *None of these answers are correct.*

Question 8 ♣ The following figures depict three different datasets, each one containing two features. In each figure, points represent the observation in the dataset, and their positions correspond to the values of the two variables. If in each of the datasets we apply Principal Component Analysis(PCA), which one(s) among the following answers is(are) correct?

Dataset 1　　　　　　　Dataset 2　　　　　　　Dataset 3

A In the third dataset the principal components could not be orthogonal.

B In the second dataset, the first principal component explains a large percentage of the total variance.

■ In the second dataset, the principal components' directions are similar to the main direction of the system of coordinates.

■ In the first dataset, the first principal component explains a large percentage of the total variance.

E In the third dataset, the first principal component explains a large percentage of the total variance.

F *None of these answers are correct.*

Question 9 ♣ The three figures depict the different outputs of the application of a Neural Network in a binary classification problem with two numerical explanatory features. All three models contain a single hidden layer but use different hyperparameter values. Green and red points correspond, in the training set, to observations in classes 0 and 1 respectively; while green and red regions show the prediction values (0 or 1) for new observations according to the three different models.
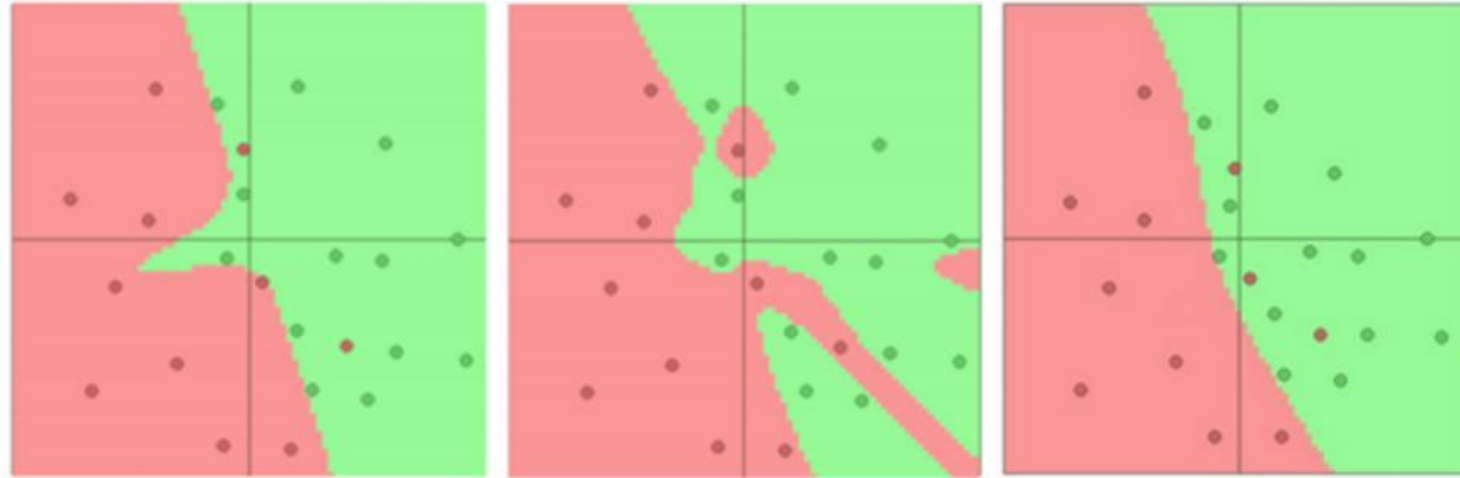
Which one(s) among the following answers is(are) correct?

A   The second model is certainly over-fitted.

B   In the third model, a better result could be obtained by modifying the regularization parameter to attribute more relevance to the regularization term.

C   It is most likely that the first model contains the largest number of neurons in the hidden layer.

D   Among the three models, it is expected that the second model will get better values for accuracy metrics on the training set.

E   Almost surely the three models use different activation functions.

F   *None of these answers are correct.*

Question 9 ♣ The three figures depict the different outputs of the application of a Neural Network in a binary classification problem with two numerical explanatory features. All three models contain a single hidden layer but use different hyperparameter values. Green and red points correspond, in the training set, to observations in classes 0 and 1 respectively; while green and red regions show the prediction values (0 or 1) for new observations according to the three different models.

Which one(s) among the following answers is(are) correct?

■ The second model is certainly over-fitted.

B In the third model, a better result could be obtained by modifying the regularization parameter to attribute more relevance to the regularization term.

C It is most likely that the first model contains the largest number of neurons in the hidden layer.

■ Among the three models, it is expected that the second model will get better values for accuracy metrics on the training set.

E Almost surely the three models use different activation functions.

F *None of these answers are correct.*

Question 10   In the text box of this question you can add any comment related to the test (withdraw, clarification, etc).