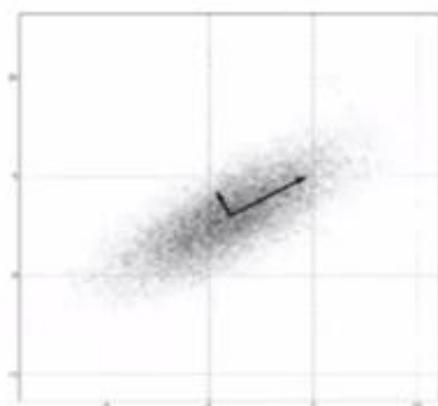


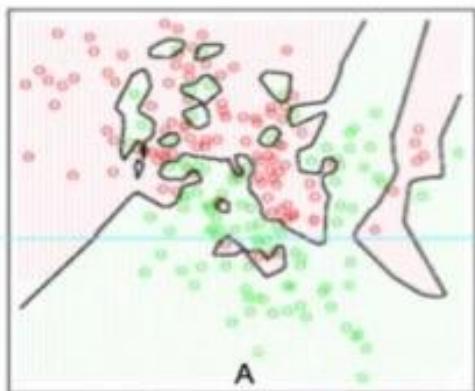
### Multiple Choice Questions

1. Which of the following sentences about Principal Component Analysis (PCA) is **false**:

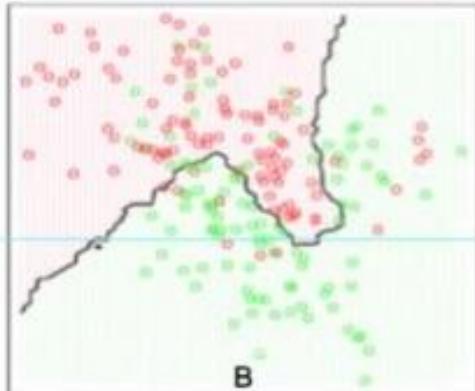
- a) In PCA, the Principal Components are linear combinations of the original variables;
- Every Principal Component accounts for an equal percentage of the total variance;
- c) If we retain all the Principal Components, then we retain all the variance in the data;
- d) None of the sentences is false;



2. In the context of the k-nearest neighbors and given the two figures (representing decision boundaries) we can say that:



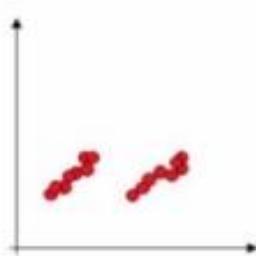
A



B

- a) Figure A uses a higher k;
- Figure B uses a higher k;
- c) Both A and B use a large k;
- d) The information provided is not enough to answer the question;

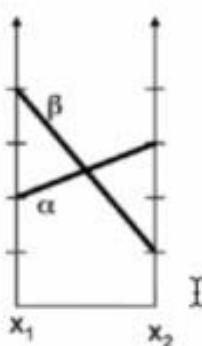
3. During classes we talked about Principal Component Analysis and its use in pre-processing tasks, particularly in reducing the input space. The figure below presents a set of data, suppose I made a PCA and I've decided to use only the 1st principal component to develop my predictive model, where the objective is to separate the two groups of points presented. I can say that:



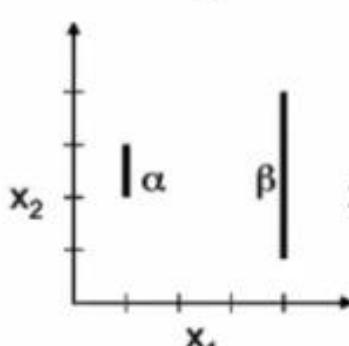
- a) the use of the PCA in this case is advisable;
- the use of the PCA in this case is not advisable;
- c) the use of the 1st principal component is particularly relevant for the purposes of the model;
- d) The information provided is not enough to answer the question;

4. Given the parallel coordinates graphic, which of the other graphics represents the same data:

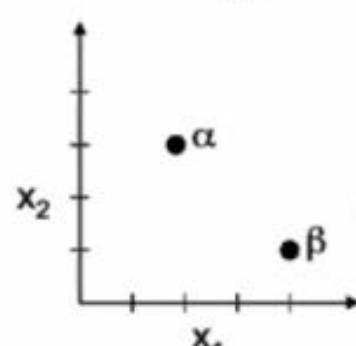
**a**



**b**



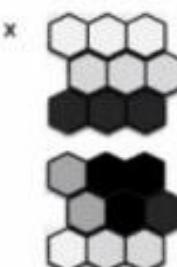
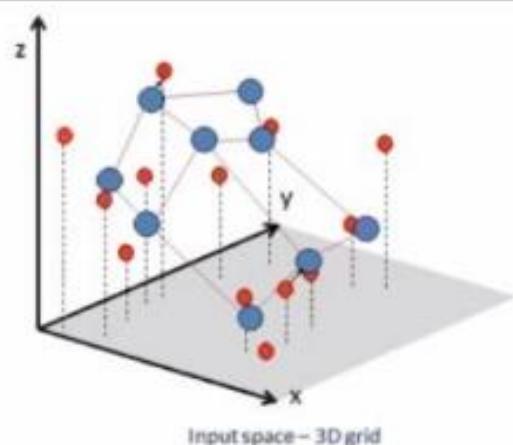
**c**



- a) Graph a;
- b) Graph b;
- c) Graph c;
- d) None of the options is correct;

5. The figure shows the input space of a SOM trained with three variables ( $x$ ,  $y$  and  $z$ ). On the right side we have 2 component planes (black = high, white = low); given that 1st component plane represents variable  $x$ , which is the variable represented in the other component plane:

- a) Variable  $y$ :
- b) Variable  $z$ ;
- c) None of the above;
- d) The information provided is not sufficient to make a decision;



6. If you have to replace the missing value representing the grade of João in the Algebra course, using the nearest neighbor ( $k=1$ ) method, the grade will be:

- a) 13;
- b) 14;
- c) 15;
- d) 16;

	SAD	Mat.	Álgebra
Manel	18	17	16
João	14	13	?
Maria	14	15	16
Pedro	10	10	11
Roberto	15	16	16
José	14	13	15
Hildefonso	13	14	13

7. During the training of a SOM, using a learning rate of 0.6, neighborhood function of 0, and input pattern  $x_1$ , and the initial weights of neurons N, shown in the table; which of the following is true?

- a) N1 will be updated to [0.76; 0.24; 0.8; 0.96];
- b) N2 will be updated to [0.04; 0.16; 0.49; 0.49];
- c) N1 will be updated to [0.92; 0.16; 0.28; 0.72];
- d) None of the options is correct;

x1	N1	N2
1	0.4	0.8
0	0.6	0.5
1	0.5	0.7
1	0.9	0.9

8. The table on the right represents the grades of students in two courses. If we standardize (normalize) SAD grades using the z-score method, José's grade will be:

- a) -1.25
- b) 0;
- c) 0.5;
- d) None of the options is correct;

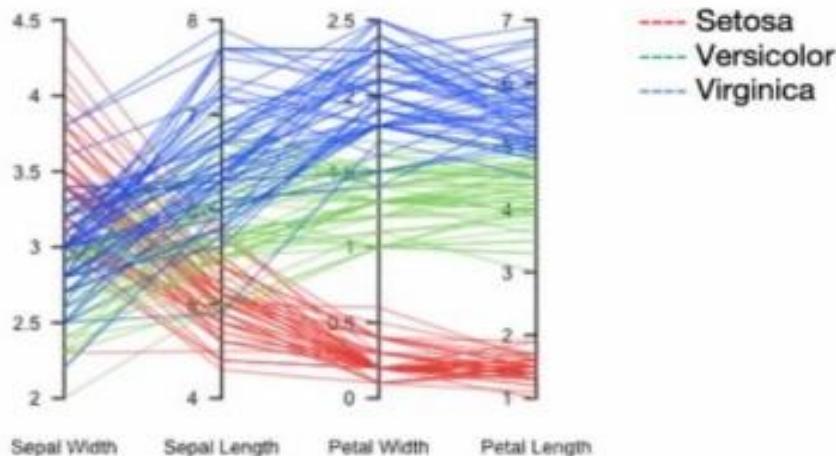
	SAD	Mat.
Manel	18	17
João	14	13
Maria	14	15
Pedro	10	10
Roberto	15	16
<b>José</b>	14	13
Hildefonso	13	14

9. Which of the following sentences is **false**:

- a) The difference between classification and regression tasks is that classification predicts a discrete label, while regression predicts a continuous quantity or value;
- b) In a SOM, in general, the learning rate decreases during training process;
- c) In k-means the number of clusters is defined before running the algorithm;
- d) None of the sentences is false;

10. Given the parallel coordinate graph shown below, we can say that, in general, the variables that measure the \_\_\_\_\_ have the highest discrimination ability between the 3 classes.

- a) Sepal;
- b) Petal;
- c) Setosa;
- d) Virginica



11. In association rules we can say:

- a) The Lift of the rule is  $X \Rightarrow Y$  is the confidence of the rule divided by the expected confidence of the consequent;
- b) The support of a rule "if X then Y" is the same as the support of the rule "if Y then X";
- c) Both options are correct;
- d) None of the options is correct;

12. Given the following table we can say that "if a customer buys X then \_\_\_ % of the times he will also buy Y"

- a) 8.5;
- b) 10;
- c) 12;
- d) None of the above;

Total Transactions	1000000
Product X	350000
Product Y	85000
Product X and Y	42000

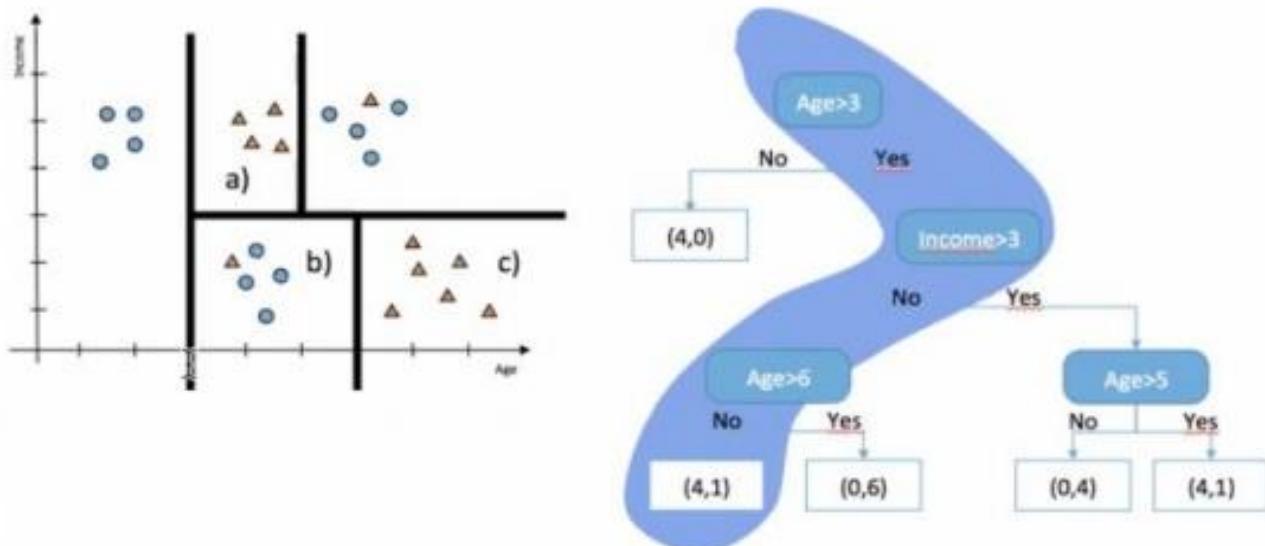
13. During the course we talked about cell-based segmentation. The situation is the following: I manage a telco company, below you can see a segmentation of my customers based on the evolution of the number of minutes used during a year. Q means quartile, dot(.) represents customers that are not part of the database in one of the periods.

Table 1

Contingency Table		Period 06.2013/06.2014					Totals
		.	<Q1	[Q1,Q2[	[Q2,Q3[	>=Q3	
Period 12.2012/ 12.2013	.	0	30603	17050	8815	6427	62895
	a	b	c	d	e	13.5%	
	<Q1	28734	55411	13600	2772	178	100695
	f	g	h	i	j	21.6%	
	[Q1,Q2[	14834	15506	52421	16838	1097	100696
	k	l	m	n	o	21.6%	
[Q2,Q3[	6540	3450	19346	59756	11608	100700	
	p	q	r	s	t	21.6%	
	3765	760	1835	14733	79604	100697	
	u	v	w	x	y	21.6%	
<b>Totals</b>		<b>53873</b>	<b>105730</b>	<b>104252</b>	<b>102914</b>	<b>98914</b>	<b>465683</b>
		<b>11.6%</b>	<b>22.7%</b>	<b>22.4%</b>	<b>22.1%</b>	<b>21.2%</b>	<b>100%</b>

- a) I can say that 79% of my best customers,  $\ge Q3$ , stayed in the same quartile between the two periods;
- b) I can say that 52% of my  $[Q1,Q2[$  customers changed quartile between the two periods;
- c) Both options are correct;
- d) None of the options is correct;

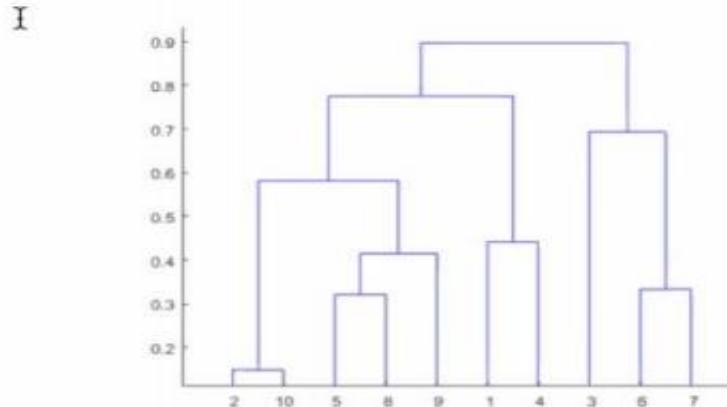
14. I've developed a predictive model using classification trees. This model allows me to classify my customers based on their propensity to buy a certain product I'm promoting. Given the tree below we can say that an individual following the path highlighted on the tree (right side) belongs to group:



- a) a);
- b);
- c) c);
- d) None of the options is correct;

15. In Hierarchical Clustering, the Dendrogram can be used to:

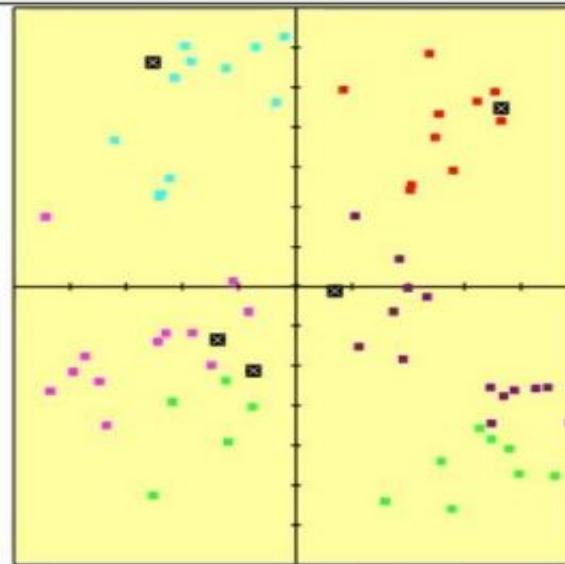
- a) Select the number of clusters as it represents the distance required by pairs of clusters to be merged
- b) Select the number of clusters as it represents the way the  $R^2$  changes with different cluster solutions
- c) Select the most adequate linkage method as it represents the distance required by pairs of clusters to be merged
- d) Select the most adequate linkage method as it represents the way the  $R^2$  changes with different cluster solutions



### True or false Questions

1. The Density-Based Clustering (DBSCAN) algorithm is able to identify outliers as noises.

2. Given the figure we can say that the optimization process of the k-means algorithm has come to an end.



3. During the course we talked about the "lie factor" concept proposed by Edward Tufte. This concept allows us to measure the relation between the size of an effect, when graphically represented, and its size in the data that originate the graphic. Let us assume that the sizes of the effect are as follows:

graphic – max: 30; min: 18  
data – max: 135000; min: 80000

Thus, we can say that, according to Tufte guidelines, the graphical effect is in between the bounds proposed as correct.

4. In the final result of a clustering, using k-means, it is possible that 2 individuals belonging to different clusters may be more similar between them than with the centroid that represents them.

5. While training a SOM with the same dataset the results may be different given different initializations.

6. The effect commonly known as the "curse of dimensionality" translates into the observation that as the number of records (observations) increases the distribution of points in space becomes sparser;