Data Mining 2

Module 1 - 2020/2021

Name ________________ **Surname** ________________ **ID:** _____________ **Test id. AUTO**

Q1. In CRISP-DM what is not done in Data Understanding phase?

1) Assess situation

2) Verify data quality

3) Determine data mining goals

4) Collect initial data

5) Describe data

A1. _______

N.B.: this question can have more than one correct answer

Q2. Given the following confusion matrix and cost matrix which is the cost of the classification?

conf matrix

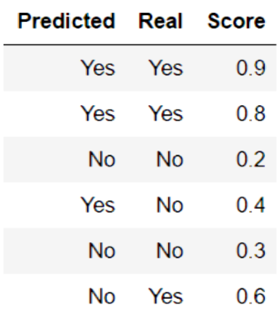
$$\begin{bmatrix} 30 & 35 \\ 5 & 30 \end{bmatrix}$$

cost matrix

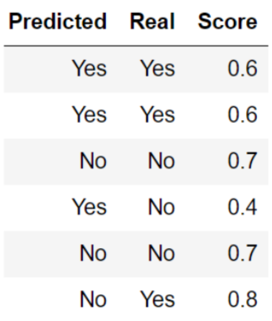
$$\begin{bmatrix} 0 & 95 \\ 5 & 0 \end{bmatrix}$$

A2. ______________________

Q3. Given the classification results in the Figure draw the corresponding Lift Charts in the plot in the Figure. If any, which result is the best predictor?



Predicted	Real	Score
Yes	Yes	0.9
Yes	Yes	0.8
No	No	0.2
Yes	No	0.4
No	No	0.3
No	Yes	0.6



Predicted	Real	Score
Yes	Yes	0.6
Yes	Yes	0.6
No	No	0.7
Yes	No	0.4
No	No	0.7
No	Yes	0.8

Cumulative
0
1
2
3
0
1
2
3
4
5
6
Cases

A3. ______________________

Q4. Which one of the following can be taken as example of imbalanced problem?

1) Iris dataset

2) Car crash

3) Disk failure

4) Rare disease

5) Soccer games results

A4. _______

N.B.: this question can have more than one correct answer

Q5. Which one of the following methods allow to deal with imbalanced problems?

1) K-NN

2) Undersampling

3) Overfitting

4) Cost sensitive classifier

5) K-Means

A5. _______

N.B.: this question can have more than one correct answer

Q6. Which is a correct description of the CNN method?

1) It is a method of class weighting

2) It is a method for undersampling

3) It is a method for oversampling

4) It is a cost sensitive classifier

5) It is an advanced version of KNN

A6. _______

Q7. Put the steps of the SMOTE algorithm in the correct order. A. Add mid-points to dataset. B. For each point get k nearest neighbors. C. Select only minority points. D. Calculate mid-points. (example of answer: A, D, C, B)

A7. ______________________

Q8. Which are the measures used for drawing the ROC plot (no matters the order)?

1) Sensitivity vs Specificity

2) TPR vs FPR

3) FPR vs FNR

4) Precision vs Recall

5) TPR vs TNR

A8. _______

Q9. Which of the following assumptions/results allow to detect an outlier using ABOD?

1) A power-law distribution of data

2) A compass-like direction of the objects around the point

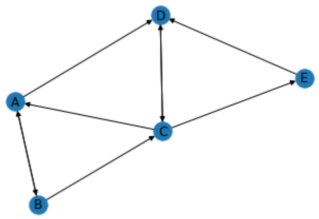
3) None of the others

4) A small variance of the angle spectrum

5) A preliminary clustering of data

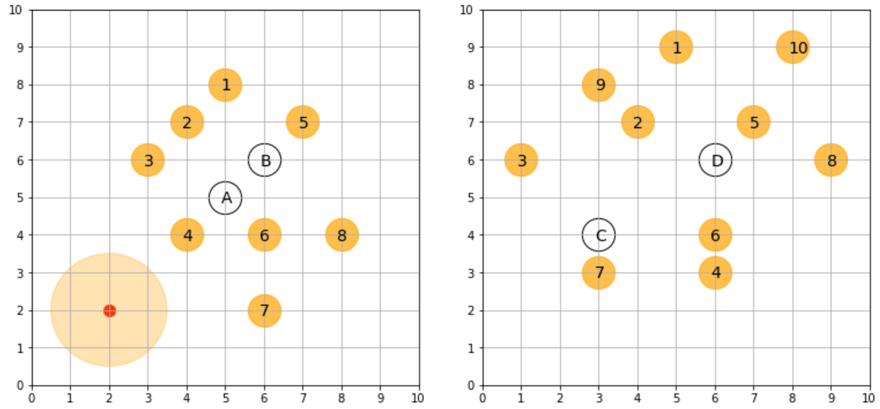
A9. _______

Q10. Given the following KNN graph induced by a set of points and a threshold $t \geq 2$, identify the outliers using the in-degrees of the nodes.



A10. _______________________

Q11. Given $\epsilon = 1.5$ and $\pi = 0.18$, are A and/or B outliers (do not count the points themselves)? Are C and/or D outliers of depth 2? Is C outlier considering the LOF of point C by taking $k = 2$? (nb: to simplify the calculus, substitute the reachability-distance with the Manhatthan distance)



A11. ______________________