15/12/2021 – IST DB2 – INSA Lyon

Name		Last Name	
Instructor	Riccardo Tommasini	Code	
Date			

You have exactly 75 minutes to complete the exam. The exam is worth 20 points (to be averaged with the project points). ***NB: Both the exam and project should do at least 12/20.***

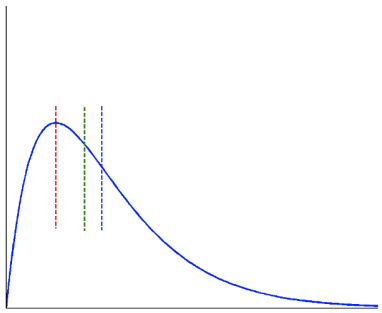
- *For multiple choice question, circle the correct answer.*
- *For open questions Do not exceed the number of lines/examples unless explicitly required. (Keep it simple)*

---

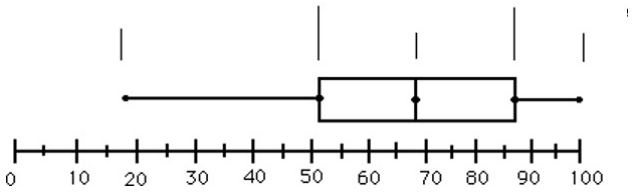
1. What is the mode of the dataset?
   a) the value that occurs most frequently in the data
   b) the value that occurs less frequently in the data
   c) the value that occurs averagely in the data

---

2. In a **positively** skewed dataset, the order relation *between mode, median and mean is….*



3. *Complete the boxplot labelling the indicated points from left to right*



4. **[Group 1]** Normalize the following group of data: 200, 300, 400, 600, 1000 using z-score normalization

   a. The normalized data are: 1.06, −0.707, −0.354, 0.354, 1.77

   b. The normalized data are: −1.06, −0.707, −0.354, 0.354, 1.77

   c. The normalized data are: −1.06, −0.707, −0.5, 0.354, 1.77

5. **[Group 2]** Normalize the following group of data: 200, 300, 400, 600, 1000 using min-max normalization by setting min = 0 and max = 1

   a. The normalized data are: 0, 0.125, 0.25, 0.5, 1
   b. The normalized data are: 0, 0.5, 0.25, 0.5, 1
   c. The normalized data are: 0, 0.5, 0.25, 0.75, 1

6. *Calculate the **cosine similarity** between two documents described by the following frequency table (report formula and calculation passages)*

Cosine:

	"awesome"	"y'all"	"basically"
Document1	3	1	5
Document2	7	3	3

7. **[Group 1]** Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215. The data is partitioned into three bins using equal-width partitioning

   a. **Bin** 1: 5, 10, 11, 13, 15, 35, 50, 55, 72 ; **Bin** 2: 92 ; **Bin** 3: 204, 215
   b. **Bin** 1: 5, 10, 11, 13, 15, 35 ; **Bin** 2: 50, 55, 72, 92 ; **Bin** 3: 204, 215
   c. **Bin** 1: 1, 5, 10, 11, 13 ; **Bin** 2: 15, 35, 50, 55 ; **Bin** 3: 72, 92, 204, 215

8. **[Group 2]** Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215. The data is partitioned into three bins using equal-frequency (equal-depth) partitioning

   a. Bin 1: 1, 5, 10, 11, 13 ; Bin 2: 15, 35, 50, 55 ; Bin 3: 72, 92, 204, 215

   b. Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72 ; Bin 2: 92 ;Bin 3: 204, 215

   c. Bin 1: 5, 10, 11, 13, 15, 35 ; Bin 2: 50, 55, 72, 92 ; Bin 3: 204, 215

9. *What of the following are **all** data preprocessing tasks*

   a. Cleaning, Reduction, Integration
   b. Cleansing, Augmentation, Join
   c. Extension, Redaction, Curation

10. **[Group 1]** List and briefly describe 3 methods do handle missing data (automatically):

11. **[Group 2]** List and briefly describe 3 methods for handling noisy data (automatically)

12. **[Group 1]** What of the following are technique for dimensionality reduction?

   a. wavelets

   b. histograms

   c. principal component analysis

   d. regression

   e. feature subset selection, creation

13. **[Group 2]** What of the following are technique for data discretization?

a. wavelets

b. histograms

c. principal component analysis

d. regression

e. feature subset selection, creation

14. *List and describe 3 differences between OLTP and OLAP*

15. Define support and confidence for association rules

16. Which of the following are algorithms for Frequent Pattern Mining?

a. Apriori

b. FPgrowth

c. Aposteriori

d. PFDecrease

17. **[Group 1]** What is overfitting? How does it affect Decision Trees, and how can be avoided?

18. **[Group 2]** What is the class imbalance problem and how can we solve it?

19. List and describe 3 differences between OLTP and OLAP

	C	-C	
C	50	30	80
-C	5	100	105
	55	130	185

20. **[Group 1]** Given the confusion matrix above calculate accuracy, error rate, and F1 (indicate formula and necessary passages)

Accuracy:

Error_rate:

F1:

21. **[Group 2]** Given the confusion matrix above, calculate sensitivity, and specificity, precision, and recall

Sensitivity:

Specificity:

Precision:

Recall:

22. **[Group 1]** List 3 characteristics of density-based clustering methods, e.g., DBSCAN

23. **[Group 2]** List 3 weaknesses of partitioning-based clustering techniques, e.g., k-means

24. Are outliers a form of noise? Elaborate your answer…

25. Comment the picture below explaining what you can infer about the used clustering techniques.