NAME	NIA	GRADE

MINING OF MASSIVE DATASETS (2022-2023)

MID-TERM EXAM

WRITE YOUR ANSWERS BRIEFLY and CLEARLY IN THE BLANK SPACES. If you do not know the answer to a question, leave it blank. No points are awarded for what does not answer the question being asked. Please underline key words in your answers. Please if you include intermediate calculations, circle the final result. If needed, you can attach an extra sheet to your exam. In this case, indicate clearly that the solution can be found in the extra sheet.

Problem 1 *1 point*

Explain **briefly** the difference between:

The data characterization and data discrimination tasks.

A nonordinal categorical attribute, and an ordinal categorical attribute.

Problem 2 *1 point*

Distribute the following data into three equi-depth and three equi-width bins: $\{10, 20, 200, 210, 300, 310, 400, 410, 1200\}$. Justify each answer.

Answer (equi-width):

Answer (equi-depth):

Problem 3 *1 point*

Define **briefly** and give an example of the following:

Range contraint:

Cross-field validation:

Problem 4 *1 point*

Define **briefly**, without using a concrete example, but a general definition.

Missing Completely at Random:

Missing at Random:

Problem 5 *1 point*

Consider the following data: $x_1 = 350, x_2 = -200, x_3 = 8, x_4 = 490, x_5 = -9500, x_6 = 0$. Perform min-max scaling and standardization, naming the min-max scaled variables $y_1, y_2, \ldots, y_6$, and the standardized variables $z_1, z_2, \ldots, z_6$. Express as a decimal number with **four digits** after the decimal point.

Min-max scaled data: $y_1 =$ $y_2 =$ $y_3 =$

$y_4 =$ $y_5 =$ $y_6 =$

Standardized data: $z_1 =$ $z_2 =$ $z_3 =$

$z_4 =$ $z_5 =$ $z_6 =$

Problem 6 *1 point*

Explain **briefly** and clearly what is the *curse of dimensionality* and what are its implications with respect to using L_p norms.

Answer:

Problem 7 *1 point*

Consider the following list of European cities:

City	Area $[km^2]$
Milan	2,225
Barcelona	1,072
Paris	2,853
Istanbul	1,471
London	1,738
Moscow	6,154
Saint Petersburg	1,510

Perform a *stratified* random sample of 6 cities, considering the following strata: less than 1,500 km^2, between 1,500 and 2,500 km^2, more than 2,500 km^2.

Answer (list of cities plus a brief explanation on how you obtained it):

Problem 8 *1 point*

Consider the shingles-document matrix below and the three given permutations.

Shingle	D1	D2
S1	0	0
S2	1	0
S3	1	1
S4	0	1
S5	1	1

π_1	π_2	π_3
1	5	4
3	1	3
2	2	5
5	4	2
4	3	1

Compute the Jaccard similarity of the sets of shingles in the two documents, expressed as a simplified fraction or as a decimal with two digits after the point. Explain your answer briefly:

Complete the signature matrix for each document:

	D1	D2
π_1		
π_2		
π_3		

Compute the similarity between the signature matrices, and express it as a simplified fraction or as a decimal with two digits after the point. Explain your answer briefly:

Problem 9 *1 point*

Consider the following database:

TID	Itemset
101	a, b, c
102	b, c
103	a, c, d, e
104	b, c, d
105	a
106	b, e

Find two closed 2-itemsets, indicate why they are closed:

Indicate the confidence of the rule $b \Rightarrow c$, expressed either as a simplified fraction or as a decimal with two digits after the decimal point. Include your intermediate calculations.

Problem 10 *1 point*

Find all frequent itemsets at minsup = 1/2 in the above database, using the *apriori* algorithm seen in class, and including tables containing your intermediate steps.

Answer: