NAME	uXXXX	GRADE

MINING OF MASSIVE DATASETS (2024-2025)

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, and when including intermediate calculations, circle the final result. If needed, you can attach an extra sheet. In this case, indicate clearly that the solution is in the extra sheet.

Problem 1

1 point

Season one of the TV series "She's Gotta Have It" (2017-2019) has the following ratings in IMDB:

Episode	S1E1	S1E2	S1E3	S1E4	S1E5	S1E6	S1E7	S1E8	S1E9	S1E10
Rating	6.9	7.1	7.3	7.3	7.5	7.1	7.4	7.3	7.7	7.6

*(a; 0.5p) If we consider this a time series, what is the **behavioral** attribute and what is the **contextual** attribute?*

*(b; 0.5p) Re-scale the ratings using **min-max scaling** (use two decimal places):*

Episode	S1E1	S1E2	S1E3	S1E4	S1E5	S1E6	S1E7	S1E8	S1E9	S1E10
Rating										

Problem 2

1 point

According to the 2023 census, the population of the 10 largest cities of Catalonia is as follows: *Barcelona: 1,655,956; L'Hospitalet de Llobregat: 276,617; Terrassa: 225,274; Badalona: 224,301; Sabadell: 217,968; Lleida: 142,990; Tarragona: 138,326; Mataró: 129,613; Santa Coloma de Gramenet: 119,195; Reus: 108,535.*

*(a; 0.5p) Categorize cities into three **equi-depth** bins:*

*(b; 0.5p) Categorize cities into three **equi-log** bins:*

Problem 3 *2 points*

Consider the following dataset:

Country	Military	Death penalty
Belize	Yes	Exists
Costa Rica	No	Abolished
El Salvador	Yes	Partially abolished
Guatemala	Yes	Partially abolished
Honduras	Yes	Abolished
Nicaragua	Yes	Abolished
Panama	No	Abolished

Show your calculations, and express your solutions below as simplified fractions, or with two decimal places.

*(a; 0.5p) Compute the **Goodall measure** between Costa Rica and Panama, considering the "Military" attribute:*

*(b; 0.5p) Compute the **Goodall measure** between Costa Rica and Panama, considering the "Death penalty" attribute:*

*(c; 1p) Compute the **Jaccard similarity** $J(A, B)$ between: (A) the set of countries that **have a military** and (B) the set of countries that have **abolished** or **partially abolished** the death penalty.*

Problem 4 *1 point*

Imagine a dataset containing meteorological variables such as windspeed, temperature, and precipitation covering multiple European cities. Provide a general **definition**, and then give an example of the following within the context of this dataset:

(a; 0.2p) Define "Mandatory Constraint"

(b; 0.3p) Give an example of what could be a Mandatory Constraint in this dataset:

(c; 0.2p) Define "Missing at Random"

(d; 0.3p) Give an example of what could be a variable Missing at Random (but not Completely at Random) in this dataset:

Problem 5 *2 points*

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

Shingle	D1	D2
S1	0	1
S2	0	0
S3	1	1
S4	0	1
S5	1	1
S6	1	0

π_1	π_2	π_3	π_4
1	3	5	4
6	2	2	1
3	1	6	2
2	4	3	3
5	5	4	5
4	6	1	6

(a; 1.5p) Write the signature matrix for each document:

	D1	D2
π_1		
π_2		
π_3		
π_4		

(b; 0.5p) Compute the similarity between the signatures of the two documents, and express it as a simplified fraction or with two decimal places. Explain your answer briefly:

Problem 6 *2 points*

Consider the following database:

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

(a; 1.5p) Find frequent itemsets at minimum support $\geq 3/5$, using the method seen in class: start with $k = 1$ and find k-itemsets for increasing k. Merge k-itemsets to form (k+1)-itemsets using the prefix rule seen in class. ***Do not create unnecessary candidate itemsets, i.e., candidates that cannot be frequent due to either the prefix merging rule or the downward closure property.*** *~~Strikethrough~~ candidates that are not frequent.*

1-itemsets	Freq.

2-itemsets	Freq.

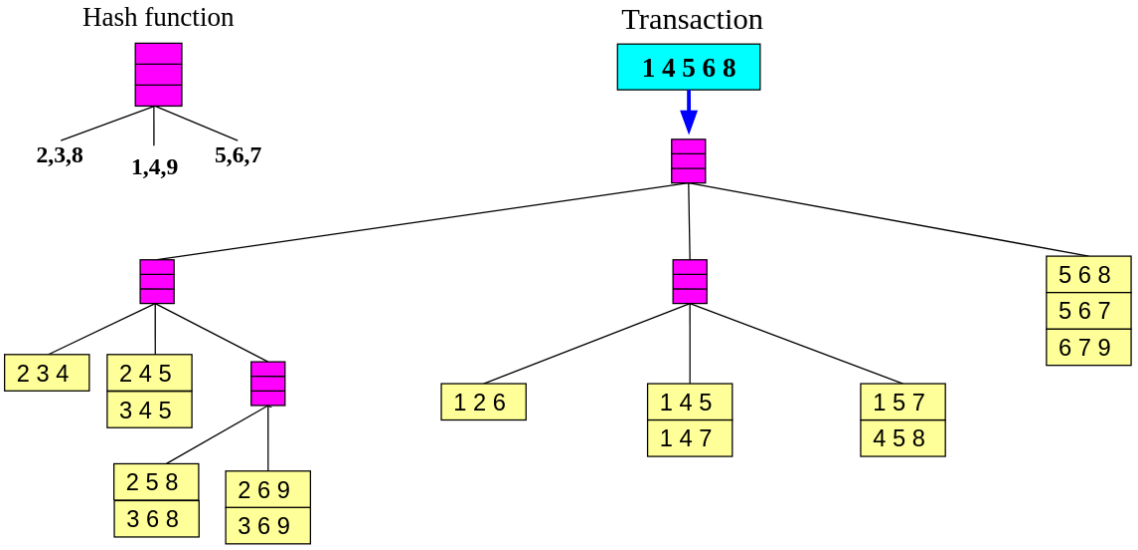
3-itemsets	Freq.

4-itemsets	Freq.

(b; 0.5p) Write one association rule with minimum support 3/5 and confidence 1:

Problem 7 *1 point*

Consider the hashtree below, and where we want to check which candidate itemsets might be contained in transaction { 1, 4, 5, 6, 8 }.



(a; 0.6p) Use the algorithm seen in class to determine which leaf nodes might contain the transaction, i.e., which are "activated" after using the algorithm seen in class. Mark them with a rectangle.

(b; 0.2p) Write the list of candidate itemsets in the activated nodes:

(c; 0.2p) Write the list of candidate itemsets that are actually contained in the transaction: