Name:

Exercise 01 (10pt): Circle the correct answer

1. Which of the following is the purpose of information retrieval (IR)?
   A. To find relevant information from a large collection of documents.
   B. To classify documents into different categories.
   C. To summarize the content of a document.
   D. To generate new documents.

2. What does a decision tree classify data points based on?
   A. The distance between points in a high-dimensional space
   B. The average value of all attributes
   C. A series of branching questions about their attributes
   D. Their proximity to known class examples

3. What is the difference between relevance and precision in IR?
   A. Relevance is a measure of how well a document matches the user's query, while precision is a measure of how many relevant documents are returned in the results.
   B. Relevance is a measure of how many relevant documents are returned in the results, while precision is a measure of how well a document matches the user's query.
   C. Relevance and precision are the same thing.
   D. Relevance is more important than precision.

4. In Apriori association rule mining, what metric tells us how often two items appear together in a transaction?
   A. Support
   B. Confidence
   C. Lift
   D. Entropy

5. A rule "Bread --> Butter" has a confidence of 70%. This means:
   A. 70% of all transactions containing bread also contain butter.
   B. 70% of transactions contain both bread and butter.
   C. 70% of transactions containing butter also contain bread.
   D. Only 30% of transactions containing bread do not contain butter.

6. How does k-means determine the optimal number of clusters?
   A. It automatically finds the best number based on data patterns.
   B. It requires the user to specify the number of clusters (k) beforehand.
   C. It uses statistical tests to identify the most significant clusters.
   D. It continues creating clusters until no further improvement is possible.

7. In the vector space model, what does each dimension of a document vector typically represent?
   A. A unique word or term from the document collection
   B. The number of paragraphs in the document
   C. The average sentence length in the document
   D. The author of the document

8. Which of the following is NOT a step in the k-means algorithm?
   A. Initializing k centroids randomly
   B. Assigning data points to their nearest centroid
   C. Updating the centroids based on the assigned points
   D. Calculating the correlation between clusters

9. Which of the following is NOT a common proximity operator used in proximity queries?
   A. NEAR
   B. WITHIN
   C. ADJ
   D. TF-IDF

10. How does a positional index typically store term positions?
   A. As a list of document IDs
   B. As a list of word offsets within documents
   C. As a hash table with term frequencies
   D. As a binary tree of term occurrence

11. What technique can robots.txt files be used for in crawling?

A. Specifying preferred crawl times
B. Blocking access to certain files
C. Providing sitemaps
D. Suggesting alternative URLs

12. How can crawlers help search engines like Google?

A. Discovering new web pages
B. Understanding the relationships between websites
C. Updating website rankings in search results
D. All of the above

13. What is the difference between supervised and unsupervised learning in data mining?

A. Supervised learning uses labeled data, while unsupervised learning uses unlabeled data.
B. Supervised learning focuses on classification, while unsupervised learning focuses on clustering.
C. Supervised learning is more accurate, while unsupervised learning is more efficient.
D. Supervised learning is for numerical data, while unsupervised learning is for categorical data.

14. We have a set of transactions containing items, and these items follow a certain pattern in their appearance within the transactions as follows:

Let T be a set of transactions and let $f(item_i)$ be the frequency of $item_i \ in \ T$. Then, for all $item_i \ in \ T$, we have:

$$f(item_i) = \sqrt{n} \ , where \ n = |T|,$$
$$the \ number \ of \ transactions \ in \ T. and \ n > 4$$

What are the frequent item sets generated by the Apriori algorithm when the support threshold is greater than or equal to 50%?

Answer: ……

Exercise 02: (6pt)

In a hospital, we want to build a decision tree for predicting the risk "Risque" of patients to have a certain disease according to their age and two boolean symptoms (V or F) called S1 and S2. The risk is evaluated according to two value

A (low), and E (high), the age is discretized according to three classes (young, adult and senior). The hospital has the data recorded in table 1.

Question: Given that the root attribute is S1 and the entropy of branch F is 0.72, determine the most suitable candidate attribute (age or S2) for this branch, including all necessary calculations.

Table 1

N°	Age	S1	S2	Risque
1	Jeune	F	V	A
2	Jeune	V	V	E
3	Adulte	F	F	A
4	Senior	F	F	E
5	Senior	F	V	A
6	Jeune	V	F	A
7	Adulte	V	F	E
8	Adulte	V	V	E
9	Senior	F	F	A
10	Senior	V	V	E