Ministry of Higher Education and Scientific Research
University of Basrah
College of Education for Pure Scienc
Department of computer

Subject: Data Mining
Stage: Fourth
Exam Time: three hours
Exam Date: 15 / 6 /2023

Final exam for the studying year 2022-2023- **First** attempt.

Notes: Answer **(*Five*)** questions, Questions in **( )** pages ,Full Marks is (50).

Q1/[marks]. (Calculator is Permission to Use)
(MCQ questions): Choose the correct answer:

1- **What are the functions of Data Mining?**
   a) Association and correlational analysis classification **b**) Prediction and characterization **c)** Cluster analysis and Evolution analysis **d**) All of the above
2- **Which of the following is an essential process in which the intelligent methods are applied to extract data patterns?**
   a) Warehousing **b**) Data Mining **c**) Text Mining **d**) Data Selection
3- **Which of the following can be considered as the correct process of Data Mining?**
   a) Infrastructure البنية التحتيه, Exploration, Analysis, Interpretation, Exploitation استغلال
   b) Exploitation , Exploration استكشاف, Infrastructure, Analysis, Interpretation
   c) Exploration, Infrastructure, Interpretation, Analysis, Exploitation **d)** Exploration, Infrastructure, Analysis, Exploitation, Interpretation
4- **Which one of the following refers to querying the unstructured textual data?**

a. Information access b) Information update c) Information retrieval d) Information manipulation

5- **Which one of the following statements about the K-means clustering is incorrect?**
   a) The goal of the k-means clustering is to partition (n) observation into (k) clusters
   b) K-means clustering can be defined as the method of quantization **c)** The nearest neighbor is the same as the K-means **d)** All of the above
6- **Suppose one wants to predict the number of newborns حديثي الولادة according to the size of storks' population by performing supervised learning**
   a) Structural equation modeling **b)** Clustering **c)** Regression **d)** Classification
7- **To classify whether incoming email is spam مزعج or not? We use:**
   a) Regression **b)** logistic regression **c)** Bays classifier **d)** none of the above
8- **Which one of the following correctly defines the term cluster?**
   a) Group of similar objects that differ significantly from other objects
   b) Symbolic representation of facts or ideas from which information can potentially be extracted **c)** Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm **d)** All of the above

9- Euclidean distance measure is can also defined as _____

a. The process of finding a solution for a problem simply by enumerating (حساب)all possible solutions according to some predefined order and then testing them

b. The distance between two points as calculated using the Pythagoras theorem (نظرية فيثاغورس) **c)** A stage of the KDD process in which new data is added to the existing selection. **d**) All of the above

10-Which one of the following can be considered as the correct application of the data mining?

a) Fraud detection (كشف الخداع) **b**) Corporate Analysis & Risk management

c) Management and market analysis **d**) All of the above

Q2/[10 marks].

A) Fill in the blanks:

1- Combing two or more attributes into a single attribute is called----------
2- Examples of data quality problems ----------
3- Benzene molecule (جزيئة البنزين)C6H6 is a type of ---------- data
4- ID numbers is a type of ---------- attributes
5- One of the advantage of data warehouse is ability to ---------- frequently

B) If the data has the following:

point	X	Y
P1	0	3
P2	3	0
P3	4	1
P4	5	1

Compute Minkowski distance if r=1; r=2; r → ∞

Q3/[10 marks].

A) 1) if p= 0 1 0 0 0 0 0 0
q= 0 0 0 0 0 1 0 1

Compute SMC and Jaccard similarity

B) Suppose you have the following data:

Transaction ID	Item bought
50	A, B, C
60	A, C
70	A, D
80	B, E, F

If min_support= 50% and min_confidence = 50%. find mining Association of rule A → C

Q4/ [10 marks].

A) Compute K-Means clustering of the following data **if K=2:**

	X	Y
A	2	2
B	3	2
C	5	4
D	6	5

B) Determine the type of attributes of the following:

a. Temperature of Kelvin
b. Eye color
c. Calendar dates
d. Ranking of pages
e. Zip codes

Q5/ [10 marks].

A Text classification is the process of assigning tags or categories to text based on their context. Assign tags(in English only) to the following data: (choose only five)

for Example : the tag of Blue, Red, Orange is Color

1-Arabic, English, Chinese, German
2- Programing in Java, Data structures, Networks
3- SSD, HD, Flash Disk
4- Sad, happy, angry, excited
5- Barcelona, Messi, FIFA world cup.
6- Grammar rules, syntax error, linker errors

B) **Consider the following data sets of patients tested for COVID-19**:

Patient	Temperature	Cough سعال	Age	Test Result
P1	*37.1*	*mild*	*young*	*positive*
P2	*37.5*	*Mild خفيف*	*Elderly كبير*	*positive*
P3	*38.2*	*Severe شديد*	*elderly*	*negative*
P4	*37.5*	*Moderate معتدل*	*middle*	*positive*
P5	*36.9*	*mild*	*elderly*	*negative*
P6	*37.8*	*Moderate*	*middle*	*negative*
P7	*37.3*	*moderate*	*young*	*negative*

If patient has Temperature =37.5, Cough= moderate and Age= elderly. Test the patient if he has positive or negative test? using Naïve Bays classifier.

Q6/ [10 marks].

A) Compute cosine similarity of the following:

d1= 2 0 3 5 0 0 2 0
d2= 1 0 0 0 1 1 0 1

B) Suppose we have a multi-class classification problem with four classes A, B, C and D. Find Precision, Recall, F1 Score of each class of Confusion Matric in Multi-Class:

		Predicted			
		A	B	C	D
Actual	A	10	2	0	1
	B	3	15	1	0
	C	1	1	8	2
	D	0	0	3	12

With my best wishes

Signature:
Examiner: *Assist.Prof . Dr. Khawla Hussein Ali*

Signature:
Head Of Department: Prof . Dr. Hamid Ali Abed Alasadi