

Name:

Enrolment No:



**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**  
End Semester Examination, December 2021

Course: Data Mining  
Program: MBA(BA)  
Course code: DSBA 7008

Semester: III  
Time: 03 hrs.  
Max. Marks: 100

**SECTION A**

1. Each Question will carry 2 Marks  
2. Instruction: Select/Write the correct answer(s)

S. No.	Question	CO
Q1.	<p>1. Hidden knowledge can be found by using _____.</p> <p>A. searching algorithm. B. pattern recognition algorithm. C. searching algorithm. D. clues.</p> <p>2. In K-nearest neighbor algorithm K stands for _____.</p> <p>A. number of neighbors that are investigated. B. number of iterations. C. number of total records. D. random number.</p> <p>3. The distance between two points that is calculated using Pythagoras theorem is _____.</p> <p>A. cartesian distance. B. euclidian distance. C. extendable distance. D. heuristic distance.</p> <p>4. Data mining algorithms require _____</p> <p>A. efficient sampling method. B. storage of intermediate results. C. capacity to handle large amounts of data. D. All of the above.</p>	CO1

5. The algorithms that are controlled by human during their execution is \_\_\_\_\_ algorithm.

- A. unsupervised.
- B. supervised.
- C. batch learning.
- D. incremental.

6. \_\_\_\_\_ analysis divides data into groups that are meaningful, useful, or both.

- A. Cluster.
- B. Association.
- C. Classification.
- D. Relation.

7. Which of the following is an extract process

- A. Capturing all of the data contained in various operational systems.
- B. Capturing a subset of the data contained in various operational systems.
- C. Capturing all of the data contained in various decision support systems.
- D. Capturing a subset of the data contained in various decision support systems.

8. Classification rules are extracted from\_\_\_\_\_.

- A. root node.
- B. decision tree.
- C. siblings.
- D. branches.

9. Which of the following is not a open source data mining tool.

- A. WEKA
- B. R
- C. RapidMiner
- D. KnowledgeMiner

10. Discovery of cross-sales opportunities are called\_\_\_\_\_.

- A. segmentation.
- B. visualization.
- C. correction.
- D. association.

**SECTION B**  
**(Scan and upload)**

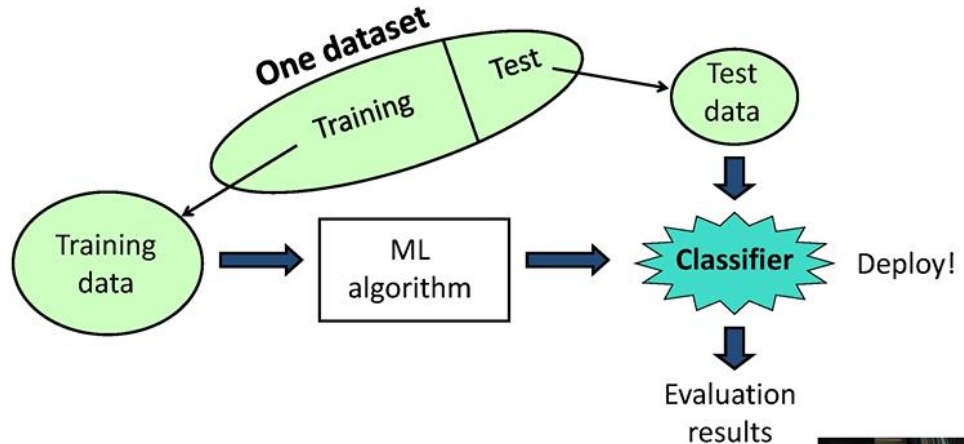
**1. Each question will carry 5 marks**

**2. Instruction: Write short/brief notes(Scan and upload)**

Q1.	<p><b>Differentiate between the following:</b></p> <ul style="list-style-type: none"> <li>a) Classification and regression</li> <li>b) Training data and Test data</li> <li>c) Cross-validation and percentage split</li> <li>d) Supervised and unsupervised learning</li> </ul>	CO2
-----	--	-----

**Section C**

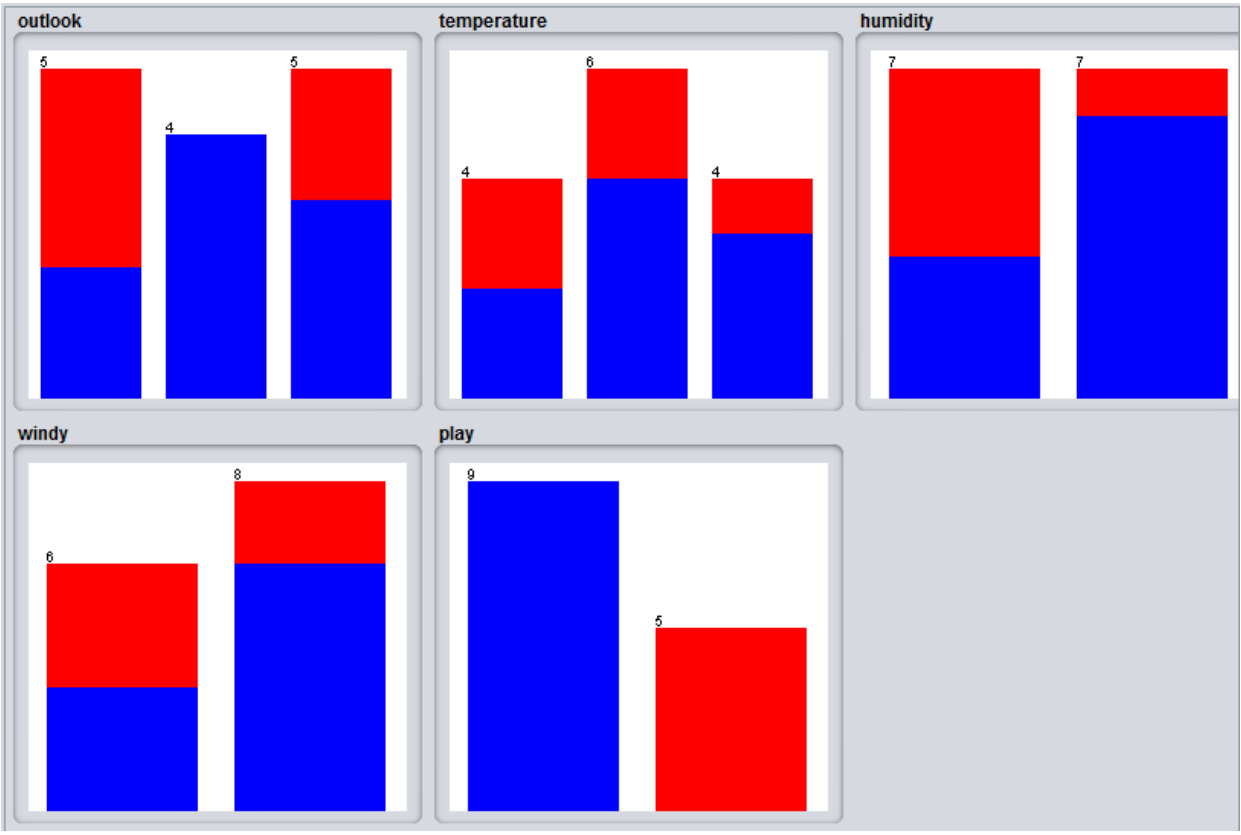
- 1. Each Question carries 10 Marks.
- 2. Instruction: Write a long answer. (Scan and upload)

Q1.	<p><b>A) Describe the process of data mining as shown in below given diagram:</b></p>  <p><b>B) Describe the below given confusion matrix:</b></p> <pre> === Confusion Matrix ===   a  b  c  d  e  f  g  &lt;-- classified as 50 15  3  0  0  1  1   a = build wind float 16 47  6  0  2  3  2   b = build wind non-float  5  5  6  0  0  1  0   c = vehic wind float  0  0  0  0  0  0  0   d = vehic wind non-float  0  2  0  0 10  0  1   e = containers  1  1  0  0  0  7  0   f = tableware  3  2  0  0  0  1 23   g = headlamps </pre>	CO2
-----	---	-----

Q2.	Based on the below data set and visualization. Write five decision rules.	CO2
-----	---	-----

**Table 1.2 The weather data.**

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

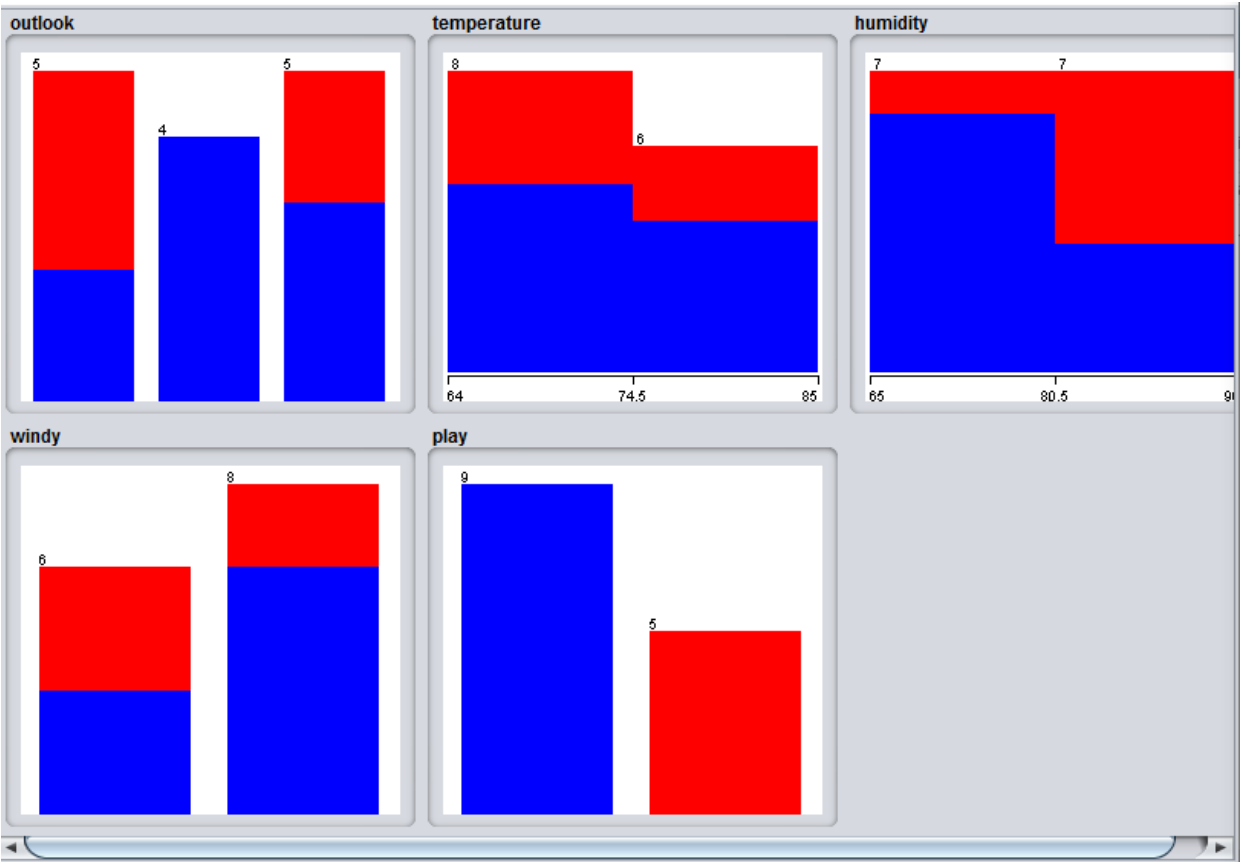


Q3. Based on the below data set and visualization. Write five decision rules.

CO2

**Table 1.3 Weather data with some numeric attributes.**

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	86	false	yes
rainy	70	96	false	yes
rainy	68	80	false	yes
rainy	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rainy	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rainy	71	91	true	no



**Section D**

1. Each Question carries 15 Marks.
2. Instruction: Write a long answer. (Scan and upload)

Q1. Considering the below data set and the result of the linear and non-linear regression analysis. Describe the interpretation of the analysis. Also, describe which analysis is better and why?

**Table 1.5      The CPU performance data.**

	Cycle time (ns) MYCT	Main memory (KB)		Cache (KB) CACH	Channels		Performance PRP
		Min. MMIN	Max. MMAX		Min. CHMIN	Max. CHMAX	
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
3	29	8000	32000	32	8	32	220
4	29	8000	32000	32	8	32	172
5	29	8000	16000	32	8	16	132
...							
207	125	2000	8000	0	2	14	52
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

**CO3**

## Linear Regression Model

class =

0.0491 \* MYCT +  
0.0152 \* MMIN +  
0.0056 \* MMAX +  
0.6298 \* CACH +  
1.4599 \* CHMAX +  
-56.075

Time taken to build model: 0 seconds

=== Cross-validation ===

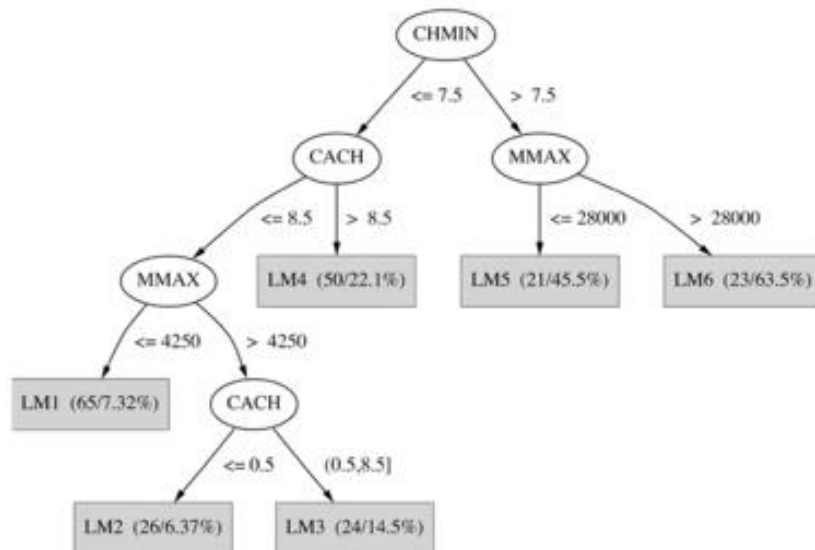
=== Summary ===

Correlation coefficient	0.9012
Mean absolute error	41.0886
Root mean squared error	69.556
Relative absolute error	42.6943 %
Root relative squared error	43.2421 %
Total Number of Instances	209

## Lesson 4.2: **NON** Linear regression

### Model tree

- ❖ Each leaf has a linear regression model
- ❖ Linear patches approximate continuous function



Q2. Describe and interpret the result of the Apriori algorithm executed on the weather file:

CO3



### Lesson 3.3: Association rules

- ❖ **Itemset** set of attribute-value pairs, e.g.

humidity = normal & windy = false & play = yes

support = 4

- ❖ 7 potential rules from this itemset:

	support	confidence
If humidity = normal & windy = false ==> play = yes	4	4/4
If humidity = normal & play = yes ==> windy = false	4	4/6
If windy = false & play = yes ==> humidity = normal	4	4/6
If humidity = normal ==> windy = false & play = yes	4	4/7
If windy = false ==> humidity = normal & play = yes	4	4/8
If play = yes ==> humidity = normal & windy = false	4	4/9
==> humidity = normal & windy = false & play = yes	4	4/14

- ❖ Generate high-support itemsets, get several rules from each
- ❖ Strategy: iteratively reduce the minimum support until the required number of rules is found with a given minimum confidence