Name: **Enrolment No:**	

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, May 2022

Course: Overview of Data Mining **Semester: II**
Program: MBA BA **Time : 03 hrs.**
Course Code: DSBA 7011 **Max. Marks: 100**

Instructions:

SECTION A
10Qx2M=20Marks

S. No.		Marks	CO1
	1. Computers are best at learning a. facts. b. concepts. c. procedures. d. principles. 2. Data used to build a data mining model. a. validation data b. training data c. test data d. hidden data 3. Supervised learning differs from unsupervised clustering in that supervised learning requires a. at least one input attribute. b. input attributes to be categorical. c. at least one output attribute. d. output attributes to be categorical. 4. Which statement is true about prediction problems? a. The output attribute must be categorical. b. The output attribute must be numeric. c. The resultant model is designed to determine future outcomes. d. The resultant model is designed to classify current behavior. 5. Which statement about outliers is true? a. Outliers should be identified and removed from a dataset. b. Outliers should be part of the training dataset but should not be present in the test data. c. Outliers should be part of the test dataset but should not be present in the training data. d. The nature of the problem determines how outliers are used.		

e. More than one of a,b,c or d is true.

6. Assume that we have a dataset containing information about 200 individuals. One hundred of these individuals have purchased life insurance. A supervised data mining session has discovered the following rule:

IF age < 30 & credit card insurance = yes
THEN life insurance = yes
Rule Accuracy: 70%
Rule Coverage: 63%

How many individuals in the class life insurance= no have credit card insurance and are less than 30 years old?

a. 63
b. 70
c. 30
d. 27

7. Unlike traditional production rules, association rules

a. allow the same variable to be an input attribute in one rule and an output attribute in another rule.
b. allow more than one input attribute in a single rule.
c. require input attributes to take on numeric values.
d. require each rule to have exactly one categorical output attribute.

8. Which of the following is a common use of unsupervised clustering?

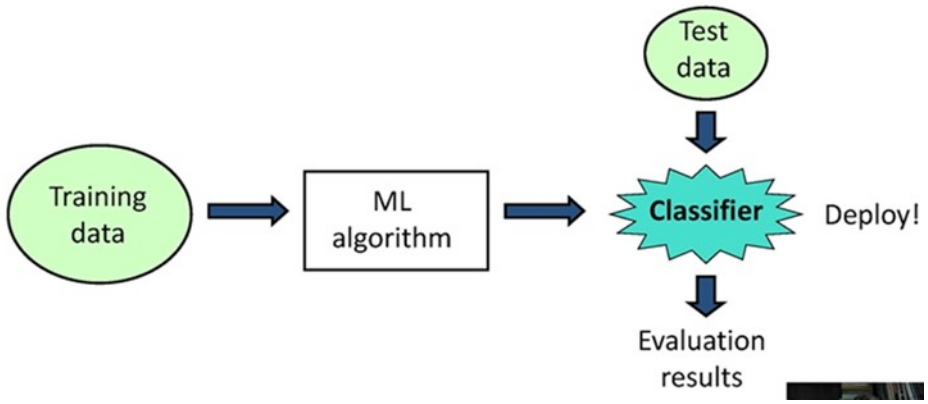
a. detect outliers
b. determine a best set of input attributes for supervised learning
c. evaluate the likely performance of a supervised learner model
d. determine if meaningful relationships can be found in a dataset
e. All of a,b,c, and d are common uses of unsupervised clustering.

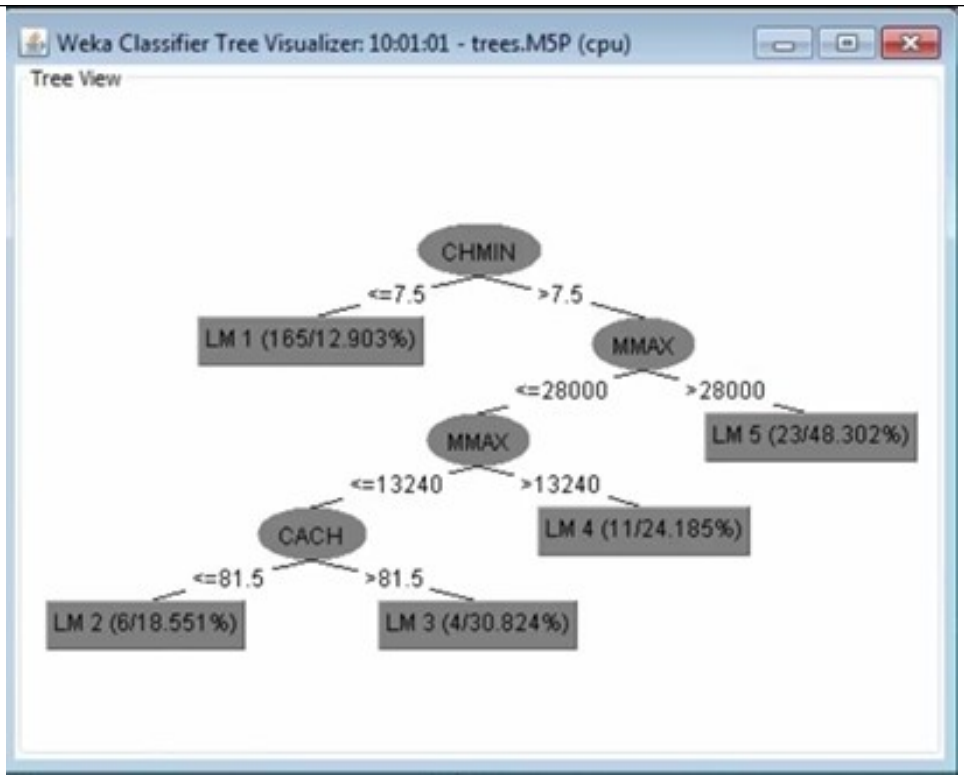
9. Which statement is true about the K-Means algorithm?

a. All attribute values must be categorical.
b. The output attribute must be categorical.
c. Attribute values may be either categorical or numeric.
d. All attributes must be numeric.

10. Classification rules are extracted from_____.

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

SECTION B **4Qx5M= 20 Marks**			
Q2.	Differentiate between training and test data set.		**CO2**
Q3.	Describe the process of evaluating J48 on any data set.		**CO2**
Q4.	Explain why cross validation better than repeated holdout.		**CO2**
Q5.	Differentiate between supervises and unsupervised learning		**CO2**
SECTION-C **3Qx10M=30 Marks**			
Q6.	Describe the below given diagram: 		**CO2**
Q7.	Write the interpretation of the association rules based on the given data set: Here are some association rules for the weather data: 1. outlook = overcast ==> play = yes 2. temperature = cool ==> humidity = normal 3. humidity = normal & windy = false ==> play = yes 4. outlook = sunny & play = no ==> humidity = high 5. outlook = sunny & humidity = high ==> play = no 6. outlook = rainy & play = yes ==> windy = false 7. outlook = rainy & windy = false ==> play = yes 8. temperature = cool & play = yes ==> humidity = normal 9. outlook = sunny & temperature = hot ==> humidity = high 10. temperature = hot & play = no ==> outlook = sunny Outlook Temp 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		**CO2**
Q8.	Describe the interpretation of below given non-linear regression result:		**CO2**



SECTION-D
2Qx15M= 30 Marks

Q9.

Table 1.5 The CPU performance data.

	Cycle time (ns) MYCT	Main memory (KB) Min. MMIN	Main memory (KB) Max. MMAX	Cache (KB) CACH	Channels Min. CHMIN	Channels Max. CHMAX	Performance PRP
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

Write the interpretation of given linear regression model based on the above data set:

PRP = -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX.

CO3

Q10. Describe the K-mean clustering algorithm with the help of example.

CO3