UNIVERSITY OF PETROLEUM & ENERGY STUDIES

DEHRADUN

End Term Examination – May, 2017

Program/course: MBA (BA)	**Semester – II**
Subject: Data Mining	**Max. Marks : 100**
Code : MBBB834	**Duration : 3 Hrs**
No. of page/s: 5	

(Please answer the questions **IN CONTEXT**)

Section - A

Q1) Select appropriate option from the following: (20 x 2 =40)

1. ..... is the process of finding a model that describes and distinguishes data classes or concepts.
A) Data Characterization
B) Data Classification
C) Data discrimination
D) Data selection

2. Classification is
A.A subdivision of a set of examples into a number of classes
B.A measure of the accuracy, of the classification of a concept that is given by a certain theory
C. The task of assigning a classification to a set of examples
D. None of these

3. Cluster is
A. Group of similar objects that differ significantly from other objects
B. Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm
C. Symbolic representation of facts or ideas from which information can potentially be extracted
D. None of these

4.Classification rules are extracted from_____.
A. root node.

B. decision tree.
C. siblings.
D. branches.

5. Which of the following is the collection of data objects that are similar to one another within the same group?
(a) Partitioning
(b) Grid
(c) Cluster
(d) Table
(e) Data source.

6. The Synonym for data mining is
(a) Data warehouse
(b) Knowledge discovery in database
(c) ETL
(d) Business intelligence
(e) OLAP.

7.Which of the following is/are the Data mining tasks?
(a) Regression
(b) Classification
(c) Clustering
(d) inference of associative rules
(e) All (a), (b), (c) and (d) above.

8. Which of the following is required by K-means clustering ?
a) defined distance metric
b) number of clusters
c) initial guess as to cluster centroids
d) All of the Mentioned
View Answer

9. Predictive analytics is same as forecasting.
a) True
b) False

10. Predicting with trees evaluate _____ within each group of data.
a) equality
b) homogeniety
c) heterogeniety
d) All of the Mentioned
View Answer

11.Computers are best at learning
a. facts.
b. concepts.
c. procedures.
d. principles.

12.Data used to build a data mining model.
a. validation data
b. training data
c. test data
d. hidden data

13. 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.

14. 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.

15. 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.

16. 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

17. 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.

18. 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.

19. 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.

20. This approach is best when we are interested in finding all possible interactions among a set of attributes.
a. decision tree
b. association rules
c. K-Means algorithm
d. genetic learning

Section – B

Attempt all questions: **(8 x 5 = 40)**

1) Differentiate between training and test data set.

2) Describe the concept of percentage split used in WEKA.

3) Describe the process of evaluating J48 on any data set.

4) Describe ZeroR classifier with the help of example.

5) Describe the 10 cross validation test option of WEKA.

6) Explain why cross validation better than repeated holdout.

7) What is cluster analysis? Give some examples of cluster analysis applications.

8) Differentiate between linear and nonlinear regression.

Section C

a) Apply k-means algorithm on below given data set (assuming k=2): (10+10)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

b) Describe the below given diagram: