DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

University Examinations 2018/2019

SEMESTER II EXAMINATIONFOR THE DEGREE OF **BACHELOR OF CRIMONOLOGY AND SECURITY STUDIES**

NAIROBI CBD CAMPUS

BSM 2416 DATA MINING

DATE: AUGUST 2018 **TIME: 2 HOURS**

Section A - Compulsory): - Select True(T) or False(F),and explain your choice on the Answer Sheet (Question 1-5 (10 points))

1) **Discriminating between spam and ham e-mails is a classification task ? [2 points]**
2) **Our use of association analysis will yield the same frequent itemsets and strong association rules whether a specific item occurs once or three times in an individual transaction ? [2 points]**
3) **The k-means clustering algorithm that we studied will automatically find the best value of *k* as part of its normal operation? [2 points]**
4) **A density-based clustering algorithm can generate non-globular clusters ? [2 points]**
5) **In association rule mining the generation of the frequent itemsets is the computational intensive step ? [2 points]**

Section B- Compulsory -: Read the following questions, Select the Correct Answer and Write it's explanation on the Answer Sheet ,From Question 6 - 15 **(40 points)**

6) Which of the following issue is considered before investing in Data Mining? **[4 points]**

A. Functionality
B. Vendor consideration
C. Compatibility
D. All of the above

7) Which of the following explanations is Classification ? **[4 points]**

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

8) Which of the following represents the explanations of Distance(Data Analysis) for Data Mining ? where *n* is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects x and y. Which of the following is a correct answer ? **[4 points]**

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

A. Binary Distance
B. Euclidean Distance
C. Mahalanobis Distance
D. Minkowski Distance

9) Which is correct answer of the following explanation ? What is Cluster ? **[4 points]**

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

10) What is the purpose of Apriori Algorithm? [4 points]

A. An influential algorithm for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties.

B. Algorithm for any mechanism employed by a learning system to constrain the search space of a hypothesis

C. An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.

D. Symbolic representation algorithm of facts or ideas from which information can potentially be extracted

11) The Key to represent relationship between tables is called? [4 points]

A. Primary key

B. Secondary Key

C. Foreign Key

D. None of these

12) Which of the following is Data independence means ? [4 points]

A. Data is defined separately and not included in programs

B. Programs are not dependent on the physical attributes of data

C. Programs are not dependent on the logical attributes of data

D. Both (B) and (C).

13) Which of the following is NOT data mining tasks that are belongs to predictive model ? [4 points]

A. Classification

B. Regression

C. Data Manipulation

D. Time series analysis

14) What is Algorithm in Data Mining ? [4 points]

A. It uses machine-learning techniques. Here program can learn from past experience and adapt themselves to new situations

B. Computational procedure that takes some value as input and produces some value as output

C. Science of making machines performs tasks that would require intelligence when performed by humans

D. None of these

15) Which is the right approach of Data Mining? [4 points]

A. Infrastructure, exploration, analysis, interpretation, exploitation

B. Infrastructure, exploration, analysis, exploitation, interpretation

C. Infrastructure, analysis, exploration, interpretation, exploitation

D. Infrastructure, analysis, exploration, exploitation, interpretation

Section C- Compulsory -: Read and Write the answers to the following questions, (From Question 16 - 20 (50 points)

16) Regarding to Data Mining , What is the difference between the following data mining tasks? Explain decisively,

a) Classification and Regression. [4 points]

b) Clustering and Association Analysis. [4 points]

17) How can you convert a decision tree into a rule set? Explain the process. [6 points]

18) List two reasons why data mining is popular now and it wasn't as popular 20 years ago. [5 points]

19) Explain how an ordinal feature differ from a nominal feature [6 points]

20) For a two-class classification problem, with a Positive class P and a negative class N, we can describe the performance of the algorithm using the following terms: TP, FP, TN, and FN.

a) What do each of these terms refer to? [10 points]
TP, TN, FP, FN

b) Place the 4 terms listed above in part a into the appropriate slots in the table below. [5 points]

		Predicted	
		Positive	Negative
Actual	Positive		
	Negative		

c) Provide the formula for accuracy in terms of TP, TN, FP, and FN. [5 points]

d) Provide the formula for precision and recall using TP, TN, FP, and FN [5 points]

--- The end of EXAM ---
Asante Sana !