



DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY

University Examination 2015/2016

**THIRD YEAR, SECOND SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR OF
BUSINESS INFORMATION TECHNOLOGY**

CBT 2301: DATA MINING

DATE: 17TH DECEMBER, 2015

TIME: 11.00AM-1.00PM

INSTRUCTIONS: Answer Question 1 and Any Other TWO Questions

Question 1

(30 Marks)

- a) Explain the following data mining terms and concepts (6 Marks)
- i) Warehouse
 - ii) Data Cube
 - iii) Decision Tree Algorithm
- b) Explain the difference between data mining, machine learning and statistics (3 Marks)
- c) Define association analysis and identify the main steps involved in the process(5 Marks)
- d) Differentiate between web structure mining and web usage mining (4 Marks)
- e) Compare the pros and cons of decision tree and neural network classification methods. (4 Marks)
- f) Explain why data collected from real world may need pre-processing and state the main tasks involved in the process (4 Marks)
- g) Outline FOUR sources of data for mining (4 Marks)

Question 2

(20 Marks)

- a) Identify FOUR popular tasks carried out during a data mining task? (4 Marks)
- b) Suppose the fraction of diploma students who smoke is 15% and the fraction of undergraduate students who smoke is 23%. If one-fifth of the college students are diploma students and the rest are undergraduates:
- i) Using the Bayesian Theorem, determine the probability that a student who smokes is a graduate student? (6 Marks)
 - ii) From the information in part (i) above, is a randomly chosen college student more likely to be a diploma or undergraduate student? Give a reason (2 Marks)

d) Discuss the components of data mining tool? (4 Marks)

e) What is the difference between OLTP and OLAP? (4 Marks)

Question 3 (20 Marks)

a) Describe the steps involved when discovering knowledge from databases. (10 Marks)

b) An Internet marketer is interested in segmenting Internet with a clustering tool using the input attributes – top ten search key words used, top 10 URLs, recent 10 online purchases (vendor, product, qty, amt), Internet usage level, heaviest access hour, and heaviest access day of a week. With a reason explain whether:

i. We can find users with different income level? (3 Marks)

ii. We can expect to find clusters differentiated based on Internet usage level? (3 Marks)

c) Discuss the main characteristics of data warehouse (4 Marks)

Question 4 (20 Marks)

a) Effective data mining needs expert knowledge about the relevant business process and knowledge about data attributes. Identify which activity requires more such expert knowledge in each of the following question. Justify your answer.

i) Configuring a clustering task versus configuring a classification task (2 Marks)

ii) Deciding on actions to be taken based on clusters discovered from a clustering task versus deciding on actions to be taken based on a decision tree. (2 Marks)

b) Discuss the role of a data warehouse and the THREE benefits a business analysts can gain from having a data warehouse? (8 Marks)

c) Describe support and confidence then express the formula for confidence using support. (4 Marks)

d) You are the network administrator of company XYZ. Explain how you would use data mining techniques to detect intrusion in the company network by hackers and other malicious software applications. (4 Marks)

Question 5 (20 Marks)

a) Define text mining and explain how it differs from data mining (3 Marks)

b) Outline THREE main techniques for web mining (3 Marks)

c) Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how each of the following data mining techniques can be applied by the company.(8 Marks)

i) clustering

ii) classification

iii) association rule mining

iv) anomaly detection

d) Discuss THREE issues associated with data mining

(6 Marks)