



MACHAKOS UNIVERSITY

University Examinations for 2019/2020 Academic Year

SCHOOL OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTING AND INFORMATION TECHNOLOGY

FOURTH YEAR SPECIAL/SUPPLEMENTARY EXAMINATION FOR

BACHELOR OF SCIENCE (COMPUTER SCIENCE)

BACHELOR OF SCIENCE (INFORMATION SCIENCE)

SIT 409: DATA MINING AND KNOWLEDGE DISCOVERY

DATE: 20/1/2021

TIME: 2:00 – 4:00 PM

INSTRUCTIONS

Answer Question One and Any Other Two Questions

QUESTION ONE (COMPULSORY) (30 MARKS)

- a) Define the following terms as used in data mining.
 - i. Classification (2 marks)
 - ii. Prediction (2 marks)
- b) What is the difference between discrete and continuous attributes? (2 marks)
- c) Discuss the Challenges of Implementing Data mining tasks (10 marks)
- d) Given a sample dataset with missing values, apply an appropriate technique of your choice to deal with them. (6 marks)
- e) What is the difference between dimensionality reduction and feature selection? (4 marks)
- f) How does a Descriptive Model differ from a Predictive Model in a data mining task? (4 marks)

QUESTION TWO S(20 MARKS)

- a) Explain "the curse of dimensionality as used in data mining and knowledge discovery (2 marks)
- b) A telecom service provider wants to increase revenues of long distance services using data mining techniques, Describe the Benefits of Data Mining (10 marks)
- c) As a system analyst of Hightech software Technologies, You note that many tuples in the company's data warehouse have no recorded values for several attributes for customer income. Explain four probable techniques that you could use to fill in these missing data. (8 marks)

QUESTION THREE (20 MARKS)

- a) Discuss the present day Applications of Data mining (10 marks)
- b) Briefly discuss the major difference between classification and clustering. List one real application for each of them. (6 marks)
- c) Differentiate between supervised and unsupervised discretization? (4 marks)

QUESTION FOUR (20 MARKS)

- a) For each data mining technique given below, describe the technique, identify which problems it is best suited for:-
- i. Decision trees (6 marks)
 - ii. Association rules (6 marks)
- b) Describe two different approaches to detect outliers in a dataset. (8 marks)

QUESTION FIVE (20 MARKS)

- a) Define sampling (2 marks)
- b) Explain any three types of sampling (3 marks)
- c) Given the following pairs of credit ranking and fraud outcome in a data set, give the formulation for entropy reduction and fill in the weights and probabilities in the formulation for splitting the records by Ranking= L. (8 marks)

	1	2	3	4	5	6	7	8	9	10
Fraud	No	No	No	Yes	No	No	No	Yes	No	Yes
Ranking	L	L	M	H	H	L	L	H	M	M

- d) Explain the factors that affect complexity data during a knowledge discovery using decision trees. (7 marks)