MACHAKOS UNIVERSITY

University Examinations for 2022/2023

SCHOOL OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTING AND INFORMATION TECHNOLOGY

FOURTH YEAR SECOND SEMESTER EXAMINATION FOR

BACHELOR OF SCIENCE (INFORMATION TECHNOLOGY)

BACHELOR OF SCIENCE (COMPUTER SCIENCE)

SIT 452 DATA MINING AND KNOWLEDGE DISCOVERY

SCO415 DATA WAREHOUSE AND DATA MINING

DATE: **TIME:**

INSTRUCTIONS

Answer Question ONE and other TWO Questions

QUESTION ONE (COMPULSORY) (30 MARKS)

a) Explain how is a data warehouse different from a database. (2 marks)

b) Data mining refers to the process or method that extracts or "mines" interesting knowledge or patterns from large amounts of data. Critics have looked at data mining as a hype that will soon go down. Using relevant examples explain the following.

i. Is data mining a hype? (4 marks)

ii. Is it a simple transformation of technology developed from databases, statistics, and machine learning? (6 marks)

c) Suppose your task as a software engineer at Machakos-University is to design a data mining system to examine the university course database, which contains the following information: the name, address, and status (e.g., undergraduate or graduate) of each student, the courses taken, and their cumulative grade point average (GPA). Describe the architecture you would choose and give the purpose of each component of this architecture. (6 marks)

d) Data quality can be assessed in terms of several issues, including accuracy, completeness, and consistency. For each of the above three issues, discuss how the assessment of data quality can depend on the intended use of the data, giving examples. (6 marks)

e) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe three various methods for handling this problem. (6 marks)

QUESTION TWO (20 MARKS)

a) Given the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20,20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

i. Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data. (6 marks)

ii. Explain ways in which you can determine outliers in the data. (4 marks)

iii. Describe TWO methods are there for data smoothing. (4 marks)

b) A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another. Give your opinion of which might be more empirically useful and state the reasons behind your answer. (6 marks)

QUESTION THREE (20 MARKS)

a) Compare two definitions (views) of prediction tasks. (2 marks)

b) Robust data loading poses a challenge in database systems because the input data are often dirty. In many cases, an input record may miss multiple values, some records could be contaminated, with some data values out of range or of a different data type than expected. Describe how you would work out an automated data cleaning and loading algorithm so that the erroneous data will be marked, and contaminated data will not be mistakenly inserted into the database during data loading. (6 marks)

c) Using real world examples discuss three data mining functionalities. (6 marks)

d) Data mining methods like attribute selection and attribute ranking will analyze the customer payment history and select important factors such as payment to income ratio, credit history, the term of the loan, etc. The results will help the banks decide its loan granting policy, and also grant loans to the customers as per factor analysis. Draw a decision tree, generate data set for the algorithm and give steps to be followed when analyzing such kind of data. (6 marks)

QUESTION FOUR (20 MARKS)

a) By use of appropriate examples discuss the following possible discoveries from a data mining exercise. (4 marks)

i. Classification

ii. Prediction

iii. Clustering

iv. Outlier Analysis

b) Consider the following confusion matrix:

a b c <– classified as

7 6 9 | a = part time

1 8 4 | b = full time

5 3 7| c = Distance learning

Use the above confusion matrix to determine the following:

(i) Precision for full time class. (2 marks)

(ii) Recall for part time class. (2 marks)

c) With the help of a diagram illustrate the Knowledge Discovery Process. (6 marks)

d) Discuss any three challenges facing data mining. (6 marks)

QUESTION FIVE (20 MARKS)

a) Using a diagram discuss how Online Analytical Mining can be integrated with Online Analytical Processing with data mining and mining knowledge in multidimensional databases. (6 marks)

b) For each data mining technique identified below, describe the technique, identify which problems it is best suited for, identify which problems it has difficulties with, and describe any issues or limitations of the technique.

i. Decision trees. (4 marks)

ii. Association rules. (4 marks)

c) The 'database' below has four transactions. What association rules can be found in this set, if the minimum support (i.e coverage) is 60% and the minimum confidence (i.e. accuracy) is 80% ? (6 marks)

Trans_id Itemlist
T1 {K, A, D, B}
T2 {D, A C, E, B}
T3 {C, A, B, E}
T4 {B, A, D}