University Examinations 2023/2024

THIRD YEAR SECOND SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR OF COMPUTER SCIENCE

CCS 3354: DATA MINING

DATE: APRIL 2024   **TIME: 2 HOURS**

INSTRUCTIONS: *Answer question **one** and any other **two** questions*

QUESTION ONE (30 MARKS)

a) The biggest challenge in data mining is to analyze data in order to extract important information that can be used to solve a problem or a company development. Using a suitable example, justify the statement.   (6 marks)

b) Data mining holds incredible potential for healthcare services due to the exponential growth in the number of electronic health records. But still, the major challenge is what should healthcare services providers do to filter all the data efficiently for improving Patient Care? How is data mining useful in such a case?   (6 marks)

c) As a system analyst of Meru Supermarket, you note that many tuples in the company's data warehouse have not recorded values for several attributes for customer income. Explain three probable techniques that you could use to fill in these missing data.

(6 marks)

d) Suppose your task as a software engineer at Meru 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. What is the purpose of each component of this architecture? (6 marks)

e) Rapid Miner is a free software with visual environment for predictive analytics designed to help data analyst. Illustrate TWO features of the software and their benefits in your data mining and analysis task at Nairobi Stock Exchange. (6 marks)

QUESTION TWO (20 MARKS)

a) Discuss Bayesian Classification as model for Data Mining. (4 marks)

i. Explain three applications of cluster Analysis. (3 marks)

ii. Explain three Requirements of Clustering in Data Mining. (3 marks)

iii. Explain three categories of Clustering methods. (3 marks)

b) What is the difference between discrimination and classification? Between characterization and clustering? For each of these pairs of tasks, how are they similar? (3 marks)

c) Differentiate between OLAP and OLTP as used in data mining. (4 marks)

QUESTION THREE (20 MARKS)

a) Using suitable examples differentiate between the following concepts

i. Data analysis and Data mining (3 marks)

ii. Descriptive Data mining and Predictive Data Mining (5 marks)

b)   Decision Tree is a supervised learning method used in data mining for classification and regression methods.

   i.   Using a suitable example discuss why they are suitable in data mining. (6 marks)

   ii.   The decision tree algorithm is based on three parameters: D, attribute list, and

           Attribute _selection method. Explain the parameters   (6 marks)

QUESTION FOUR (20 MARKS)

a)   Data warehouse is "an integrated, subject-oriented, time-variant, nonvolatile collection of data that provides support for decision making.". Expound in relationship to Meru Teachers Sacco.   (8 marks)

b)   Given five frequent itemsets: {beach, ocean}, {beach, sunshine}, {holiday, sunshine}

     {ocean sunshine} and {beach, ocean sunshine}.

     (Min_Confidence =70;Confidence$(X \rightarrow Y) = P(X/Y) = P(XuY)/P(X)$)

       i. Generate association rules from the frequent itemsets.   (2 marks)

       ii. Calculate the confidence of each rule   (2 marks)

       iii.   Identify all the strong association rules.   (2 marks)

c)   Differentiate between the following feature transformation techniques

       i. Log Transformation   (3 marks)

       ii. Reciprocal transformation   (3 marks)

QUESTION FIVE (20 MARKS)

a) The CRISP-DM (cross-industry standard process) methodology provides a structured
    approach to planning a data mining project.

i. Discuss the concept of CRISP and how it helps in data mining.   (4 marks)

ii. Using a suitable example explain the following two phases in CRISP

      1.   Business Understanding   (8 marks)

      2.   Data Understanding   (8 marks)