Kasarani Campus
Off Thika Road
P. O. Box 49274, 00101
NAIROBI
Westlands Campus
Pamstech House
Woodvale Grove
Tel. 4442212
Fax: 4444175

# KIRIRI WOMENS' UNIVERSITY OF SCIENCE AND TECHNOLOGY
# UNIVERSITY EXAMINATIONS, 2024/2025 ACADEMIC YEAR
# FOURTH YEAR, SECOND SEMESTER EXAMINATION
# FOR THE DEGREE OF BACHELOR OF SCIENCE (COMPUTER SCIENCE)

## KCS 2415: DATA MINING

**DATE: 10TH DECEMBER, 2024**
**TIME: 11:30AM-1:30PM**

**INSTRUCTIONS TO CANDIDATES**
**ANSWER QUESTION ONE (COMPULSORY) AND ANY OTHER TWO QUESTIONS**

## QUESTION ONE: COMPULSORY (30 MARKS)

a) According to Data Never sleeps 9.0, TikTok users spent around 840 MegaBytes (MBs) per hour on the app in 2023. Convert the MBs to bits. **(4 Marks)**

b) Looking at the descriptions, as well as the feature values in the table below, list the data Types according to their features.

| ID | Genre | Rating | Gross | Cinema |
|---|---|---|---|---|
| 1 | Horror | Very Bad | 5000 | 0 |
| 2 | Drama | Good | 8000 | 1 |
| 3 | Comedy | Very Bad | 9000 | 1 |

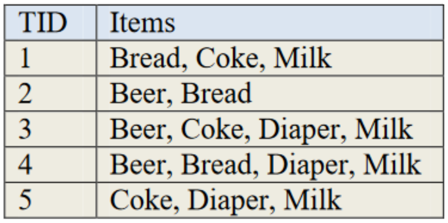
**genre**: Contains various names of movie styles. **rating:** Movies are rated on a 5 point scale from very bad to very good. **gross**: Money that the movie made. **cinema:** If the movie was shown in cinemas.

i) Highlight the nominal data types **(2 Marks)**

ii) Name the ordinary data types **(2 Marks)**

iii) List the binary data types **(2 Marks)**

c) Given the following set of records produce dependency rules which will predict occurrence of an item based on occurrences of other items. **(6 Marks)**



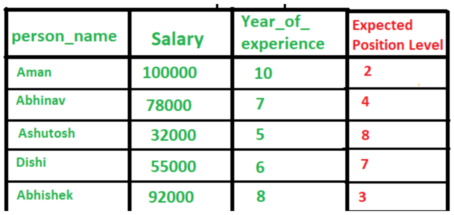
| TID | Items |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

d) Say you need to distribute 100 balls over 5 boxes.

i) Explain in what situation the entropy of the distribution is the highest. **(4 Marks)**

ii) Calculate the mathematical maximum entropy of the balls and the boxes. **(5 Marks)**

e) Given these sample datasets, use any data transformation technique to normalize only the *Expected Position Level* Column. **(5 Marks)**

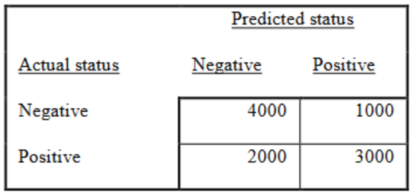


| person_name | Salary | Year_of_ experience | Expected Position Level |
|---|---|---|---|
| Aman | 100000 | 10 | 2 |
| Abhinav | 78000 | 7 | 4 |
| Ashutosh | 32000 | 5 | 8 |
| Dishi | 55000 | 6 | 7 |
| Abhishek | 92000 | 8 | 3 |

## QUESTION TWO: (20 MARKS)

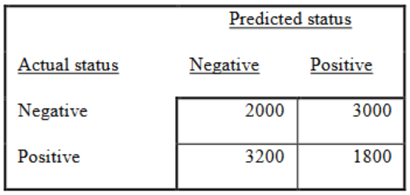
a) A team of researchers at CDC-Kenya would like to develop a predictive algorithm for TB in Kenya using two explanatory variables: HIV status; and smoking status.

**Fit 1: HIV status as the only predictor**



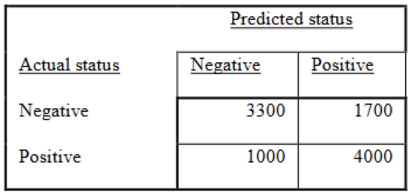
| | Predicted status | |
|---|---|---|
| Actual status | Negative | Positive |
| Negative | 4000 | 1000 |
| Positive | 2000 | 3000 |

**Fit 2: smoking status as the only predictor**



| | Predicted status | |
|---|---|---|
| Actual status | Negative | Positive |
| Negative | 2000 | 3000 |
| Positive | 3200 | 1800 |

**Fit 3: HIV and smoking predictors**



| | Predicted status | |
|---|---|---|
| Actual status | Negative | Positive |
| Negative | 3300 | 1700 |
| Positive | 1000 | 4000 |

i) From the confusion matrices above, compare the three data mining algorithms fit. Compare your results based on model accuracy. **(4 Marks)**

ii) For the best fitting data mining algorithm, compute the following measures: sensitivity, specificity and the false positive rate. **(6 Marks)**

b) Using Occam's Razor principality to choose the best hypothesis in the data below. Briefly describe the principle in context with the given data and chosen hypothesis. **(5 Marks)**

• Data:

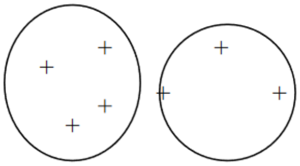
E = {0, 000, 00000, 0000000, 000000000}

• Hypotheses:

G1 : S → 0|000|00000|0000000|000000000

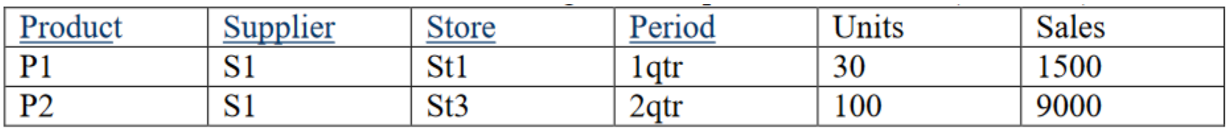
G2 : S → 00S|0

c) The two clusters shown below are well separated. Answer with either TRUE or FALSE. Justify your answer. **(5 Marks)**



## QUESTION THREE:(20 MARKS)

a) Data is organized around one or more fact tables. Each Fact Table collects a set of homogeneous events (facts) characterized by dimensions and dependent attributes.

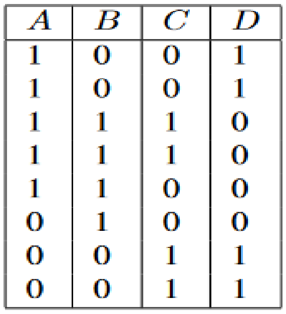


| Product | Supplier | Store | Period | Units | Sales |
|---|---|---|---|---|---|
| P1 | S1 | St1 | 1qtr | 30 | 1500 |
| P2 | S1 | St3 | 2qtr | 100 | 9000 |

i) With the aid of the data in the table above design a conceptual star schema. **(5 Marks)**

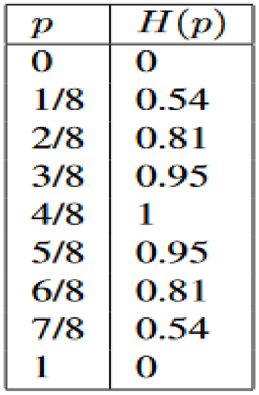
ii) With the aid of the data in the table above design a snowflake schema schema. **(5 Marks)**

b) Consider the following dataset of binary attributes (Story B: Maximally Informative k-Itemsets):



| $A$ | $B$ | $C$ | $D$ |
|---|---|---|---|
| 1 | 0 | 0 | 1 |
| 1 | 0 | 0 | 1 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 1 | 0 |
| 1 | 1 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 1 |

For answering the following questions, you may need to consult the following table of entropies.



| $p$ | $H(p)$ |
|---|---|
| 0 | 0 |
| 1/8 | 0.54 |
| 2/8 | 0.81 |
| 3/8 | 0.95 |
| 4/8 | 1 |
| 5/8 | 0.95 |
| 6/8 | 0.81 |
| 7/8 | 0.54 |
| 1 | 0 |

i) Give the entropy of each of the four attributes over the entire dataset of Story B. **(4 Marks)**

ii) List the itemset(s) having a miki with k = 2. **(4 Marks)**

iii) Give the joint entropy of {A, B, C, D} **(2 Marks)**

## QUESTION FOUR:(20 MARKS)

A large e-commerce company wants to better understand its customers by grouping them into segments based on their shopping behavior. Traditional clustering algorithms like k-means often require predefined clusters and struggle with complex, nonlinear relationships in the data. Therefore, the company decides to use a genetic algorithm (GA) to perform customer segmentation as a data mining approach.

The genetic algorithm works by encoding potential clustering solutions as chromosomes, where each gene represents a customer assigned to a particular cluster. The fitness function evaluates each solution based on intra-cluster similarity (how similar customers within the same group are) and inter-cluster dissimilarity (how different the groups are from each other).

Over generations, the GA applies selection, crossover, and mutation to evolve better clustering solutions. The final result is a set of customer clusters with high similarity within each group and significant dissimilarity between groups.

a) Describe why a genetic algorithm might be more effective for customer segmentation in this case study than traditional clustering methods such as k-means. **(5 Marks)**

b) Explain the role of the fitness function in the genetic algorithm used in this case study for customer segmentation. **(5 Marks)**

c) Discuss the purpose of using crossover and mutation in the genetic algorithm for this customer segmentation task **(4 Marks)**

d) Identify two limitations of using a genetic algorithm for clustering tasks like customer segmentation and suggest a potential solution. **(6 Marks)**

## QUESTION FIVE:(20 MARKS)

A hospital network aims to improve early detection of heart disease by analyzing patient data with a deep neural network (DNN). Doctors face challenges in predicting heart disease because many patients show overlapping symptoms with other conditions, and disease onset often varies widely across age, lifestyle, and genetic factors.

The hospital gathers a dataset including patient age, blood pressure, cholesterol levels, heart rate, and family history, along with previous diagnoses and lifestyle factors such as diet and exercise. The DNN model is trained to classify whether a patient has a high or low risk of developing heart disease. Given the sequential nature of medical visits, a recurrent neural network (RNN) is added to capture the history of each patient's data over time.

To handle class imbalance—since most patients do not have heart disease—and to account for noisy data, such as varying cholesterol levels from visit to visit, the hospital team employs several techniques to enhance model accuracy. Despite these adjustments, the team faces challenges with interpretability, needing to explain the model's predictions to medical staff and patients.

a) Explain why a deep neural network (DNN) could be more effective than conventional models for diagnosing heart disease in this scenario. **(5 Marks)**

b) Discuss two strategies to address the issue of class imbalance in this heart disease prediction model. **(6 Marks)**

c) List some potential impacts of noisy data, such as fluctuating cholesterol levels, on the neural network's performance? Suggest a method to minimize these effects. **(6 Marks)**

d) Given the interpretability requirements, briefly suggest and explain one method to help explain the model's predictions to medical professionals and patients. **(3 Marks)**