INTERNATIONAL UNIVERSITY OF EAST AFRICA
'Learning to Succeed'

FACULTY OF SCIENCE AND TECHNOLOGY

END OF SEMESTER EXAMINATIONS - APRIL 2025

PROGRAMME: BSSE
YEAR/SEM: YEAR 4/SEMESTER 2
COURSE CODE: SWE4202
NAME: DATA MINING AND WAREHOUSING

DATE:
TIME:

INSTRUCTIONS TO CANDIDATES:

SECTION A

This section is compulsory

Question 1:

The architecture of data mining involves several key components that work together to extract valuable information from large datasets. These components include the user interface, knowledge base, pattern evaluation, data mining engine, and data warehouse. Answer the following questions based on your understanding of these components and their roles in data mining:

a. a. Describe the architecture of data mining. (6 marks)

b. b. Explain the role of the following components in the data mining architecture:

i. i. User Interface (6 marks)

ii. ii. Knowledge Base (6 marks)

iii. iii. Data Mining Engine (8 marks)

iv. iv. Pattern Evaluation (6 marks)

v. v. Data Warehouse (8 marks)

Question 2:

You are given a dataset collected from multiple sources for a business intelligence project. The dataset contains missing values, duplicate records, outliers, and inconsistent formats.

(a) Explain four major tasks in data preprocessing and how they address data quality issues. (24 Marks)
(b) Discuss the importance of data preprocessing in data mining and data warehousing. (8 Marks)
(c) Describe two real-world scenarios where data preprocessing plays a critical role. (8 Marks)

SECTION B

Answer any three questions from this section

Question 1:

a. Explain the term data mining. (4 marks)

b. Explain the major steps involved in the data mining process. (6 marks)

c. Why do we need data mining? (6marks)

d. Explain the key properties of data mining (4marks)

Question 2:

a. What is the difference between supervised and unsupervised learning in data mining? Provide examples of each. (3 marks)

b. Explain at least four areas where data mining can be applied (8 marks)

c. Discuss at least three types of data mining techniques and give an example of how each technique is used. (9 marks)

Question 3:

a. Explain the six steps involved in data mining process (12 marks)

b. Explain the issues that are involved in data mining (8 marks)

Question 4:

(a) Explain the key steps involved in the Apriori algorithm. (6 Marks)

(b) Define support, confidence, and lift, and explain their significance in association rule mining. (6 Marks)

(c) Given the following transactions, determine the frequent itemsets using a minimum support of 50%. (4 Marks)

Transaction ID	Items Purchased
1	Milk, Bread, Butter
2	2 | Milk, Bread

3	Bread, Butter
4	Milk, Butter
5	Milk, Bread, Butter

d. Identify one association rule from the frequent itemsets and calculate its confidence. (4 Marks)

Question 5:

A supermarket chain wants to analyze customer purchasing behavior using association rule mining. The transactions database contains records of items bought together by different customers.

a) Explain the concept of association rule mining and its importance in data mining and warehousing. (5 Marks)

b) Define the following key measures used in association rule mining:

   a. Support

   b. Confidence

   c. Lift (5 Marks)

c) Given the following transaction dataset:

Transaction ID	Items Purchased
T1	Bread, Butter, Milk
T2	Bread, Diapers, Beer

T3	Bread, Butter, Diapers
T4	Milk, Diapers, Beer
T5	Bread, Milk, Diapers, Beer

If the minimum support threshold is 40% and the minimum confidence threshold is 50%, identify whether the rule {Bread} → {Diapers} is a strong association rule. (5 Marks)

a. Describe one real-world application of association rule mining in a business environment. (5 Marks)

Question 6:

a. Explain the concept of a data warehouse and its significance in business intelligence. (5 Marks)

b) Describe the four key building blocks of a data warehouse:

i. Data sources

ii. Data staging

iii. Data storage

iv. Data presentation (5 Marks)

c) Differentiate between OLAP (online analytical processing) and OLTP (online transaction processing) in the context of data warehousing. (5 Marks)

d) Discuss the role of ETL (Extract, Transform, Load) in building a data warehouse and explain its three main stages. (5 Marks)

Question 7:

a) Define data mining and explain its significance in modern data analysis. (4 Marks)

b) Discuss the motivation behind data mining and identify two major challenges faced in its implementation. (4 Marks)

c) Explain the key tasks of data mining and provide an example ofeach. (4 Marks)

d) Differentiate between structured and unstructured data and explain two common types of datasets used in data mining. (4 Marks)

e) **Describe two real-world applications of data mining and explain how they benefit organizations. (4 Marks)**

Question 8:

a) Explain the concept of data quality and discuss three key characteristics of high-quality data. (4 Marks)

b) Describe the role of data preprocessing in data mining and explain why it is necessary. (4 Marks)

c) Briefly explain the following data preprocessing techniques and provide one example for each:

i. Aggregation

ii. Sampling

iii. Dimensionality reduction

iv. Feature selection (4 Marks)

d) Define discretization and transformation in data preprocessing, and explain how they help improve data mining efficiency. (4 Marks)

e) Explain how similarity and dissimilarity between data objects can be measured and describe one technique used for each. (4 Marks)

Question 9:

a) Explain the concept of a decision tree and its importance in classification. (4 Marks)

b) Describe the ID3 algorithm and explain how it constructs a decision tree. (4 Marks)

c) Explain the improvements made in the C4.5 algorithm

compared to ID3. (4 Marks)

d) Define rule induction and describe the working mechanism of the RIPPER algorithm. (4 Marks)

e) Compare decision trees and rule induction by highlighting two advantages and two disadvantages of each approach. (4 Marks)

Question 10:

a) Explain the concept of rule-based classification and discuss its advantages over decision tree classification. (4 Marks)

b) Define associative classification and compare the CBA and MMAC approaches in terms of their working principles. (4 Marks)

c) What is rule pruning? Explain how the Reduced Error Pruning (REP) and Database Coverage methods help improve classification accuracy. (4 Marks)

d) Discuss the role of regression, neural networks, and genetic algorithms in classification. Provide one example where each technique is applied. (4 Marks)

e) Explain why classification is an essential data mining task and describe one real-world application where rule-based classification is beneficial. (4 Marks)

Question 11:

A retail company wants to analyze transaction data to discover frequent itemsets efficiently. They decide to use the Frequent Pattern Growth (FP-Growth) algorithm instead of Apriori due to its efficiency.

a. a. Explain the main steps of the FP-Growth algorithm and how it differs from the Apriori algorithm. *(5 marks)*

b. b. Describe the structure and purpose of an FP-Tree in the FP-Growth algorithm. *(5 marks)*

Transaction ID	Items Purchased

T1	Milk, Bread, Butter
T2	Bread, Butter, Jam
T3	Milk, Bread
T4	Bread, Butter
T5	Milk, Butter

c. Consider the transaction dataset given above

d.

c. Construct an FP-Tree for the given dataset, assuming a minimum support of 2. *(5 marks)*

d. d. Using the FP-Tree, extract the frequent itemsets using the FP-Growth algorithm. *(5 marks)*

Question 12:

a. Explain what data mining is and describe its key objectives in the context of business intelligence. *(4 marks)*

b. Discuss two common data mining techniques (e.g., classification, clustering, association rule mining) and provide a brief example of how each technique might be applied in a real-world scenario. *(6 marks)*

c. Define a data warehouse and explain its main components. How does a data warehouse differ from a traditional operational database? *(5 marks)*

d. Describe the role of the ETL (Extract, Transform, Load) process in data warehousing. Why is data integration critical in a data warehouse environment? *(5 marks)*