UNIRAZAK
UNIVERSITI TUN ABDUL RAZAK

FINAL EXAMINATION

MARCH 2024

COURSE TITLE	**DATA MINING**
COURSE CODE	**RBAN3213**
DATE/DAY	**22 JUNE 2024 / SATURDAY**
TIME/DURATION	**05:00 PM - 07:00 PM / 02 Hour(s) 00 Minute(s)**

INSTRUCTIONS TO CANDIDATES :

1. Please read the instruction under each section carefully.
2. Candidates are reminded not to bring into examination hall/room any form of written materials or electronic gadget except for stationery that is permitted by the Invigilator.
3. Students who are caught breaching the Examination Rules and Regulation will be charged with an academic dishonesty and if found guilty of the offence, the maximum penalty is expulsion from the University.

(This Question Paper consists of **9** Printed Pages including front page)

*****DO NOT OPEN THE QUESTION PAPER UNTIL YOU ARE TOLD TO DO SO*****

This question paper consists of TWO (2) sections. Answer ALL questions in the answer booklet provided. **[100 MARKS]**

SECTION A **(40 Marks)**

There are TWENTY (20) questions in this section. Answer ALL questions in the answer booklet.

1. Which of the following is an essential process in which intelligent methods are applied to extract data patterns?

   A. Warehousing
   B. Text Mining
   C. Data Selection
   D. Data Mining

2. Which of the following refers to the problem of finding abstracted patterns (or structures) in the unlabeled data?

   A. Supervised Learning
   B. Unsupervised Learning
   C. Hybrid Learning
   D. Reinforcement Learning

3. What are the functions of Data Mining?

   A. Association and Correctional Analysis Classification
   B. Prediction and Characterization
   C. Cluster Analysis and Evolution Analysis
   D. All of the above

4. The self-organizing maps can also be considered as the instance of _____ type of learning.

   A. Supervised Learning
   B. Unsupervised Learning
   C. Missing Data Amputation
   D. Both A & C

5. Suppose one wants to predict the number of newborns according to the size of storks' population by performing supervised learning.

   The statement above can be considered as the example of_____

   A. Classification
   B. Regression
   C. Clustering
   D. Structure equation modelling

6. Which of the following is not an issue in data mining?

   A. High dimensionality
   B. Shortage of Data
   C. Outliers
   D. Overfitting

7. _____ are any facts, numbers, images, or text that can be processed by a computer.

   A. Information
   B. Data
   C. Instructions
   D. Codes

8. Business Intelligence and data warehousing is used for _____

   A. Forecasting
   B. Data Mining
   C. Both A & B
   D. None of the above

9. Which technique can be used to predict the number of newborn babies based on existing data?

   A. Clustering
   B. Classification
   C. Regression
   D. Outlier Detection

10. Assume that you have trained a linear regression model to predict housing prices in your neighbourhood. Which function do you check to determine the proportion of samples that your model has predicted wrongly?

    A. Linear Function
    B. Activation Function
    C. Loss Function
    D. Objective Function

11. Which of the following is not a data mining application?
    A. Fraud Detection
    B. Image Recognition
    C. Customer Segmentation
    D. Speech Recognition

12. A telecommunication company wants to segment their customers into distinct groups in order to send appropriate subscription offers. This is an example of

    A. Unsupervised learning
    B. Supervised learning
    C. Knowledge representation
    D. Data transformation

13. Which design pattern would you use? Which of the following is not a data mining functionality?

    A. Characterization & Discrimination
    B. Classification & Regression
    C. Selection & Interpretation
    D. Clustering & Analysis

14. Mark the following sentiments as either implicit or explicit:

    Example I: "The software update is so bad that everything is slow on my phone now."

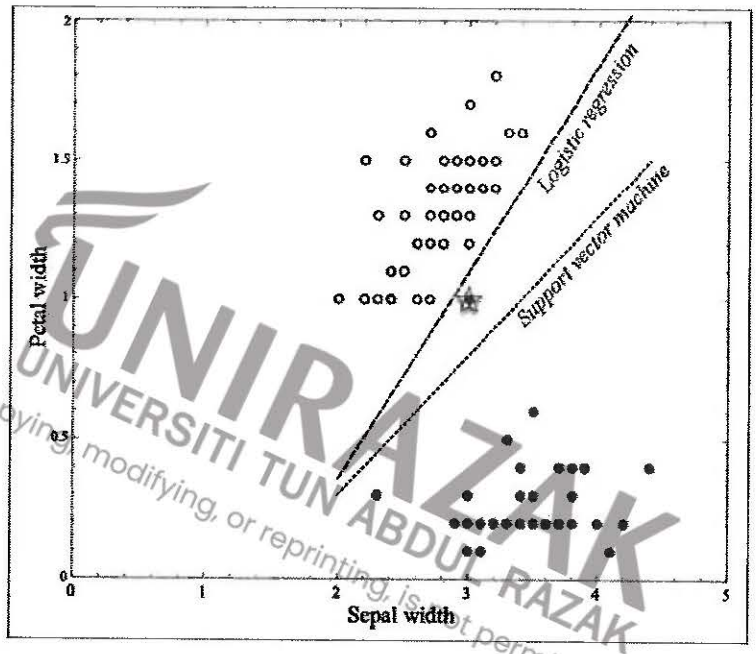
    Example II: "Software update again? You think I have so much time is it?"

    A. Example I: Implicit, Example II: Explicit
    B. Example I: Implicit, Example II: Implicit
    C. Example I: Explicit, Example II: Explicit
    D. Example I: Explicit, Example II: Implicit

15. A common method used by some data mining techniques to deal with missing data items during the learning process.

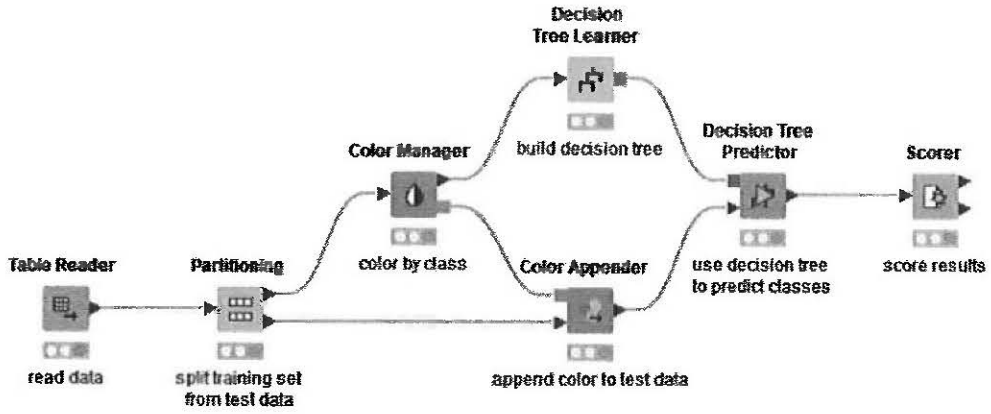
A. Replace missing real-valued data items with class means.
B. Discard records with missing data
C. Replace missing attribute values with the values found within other similar instances
D. Ignore missing attributes

16. Assume that you have implemented two machine learning algorithms in categorizing Iris flowers based on two features: sepal width and petal width. You plot the decision boundaries as shown in the figure above, one of logistic regression and another one of support vector machines. When you present the outcome to your team, how will you explain it when your boss asked: "Why are the decision boundaries of the two algorithms different?"



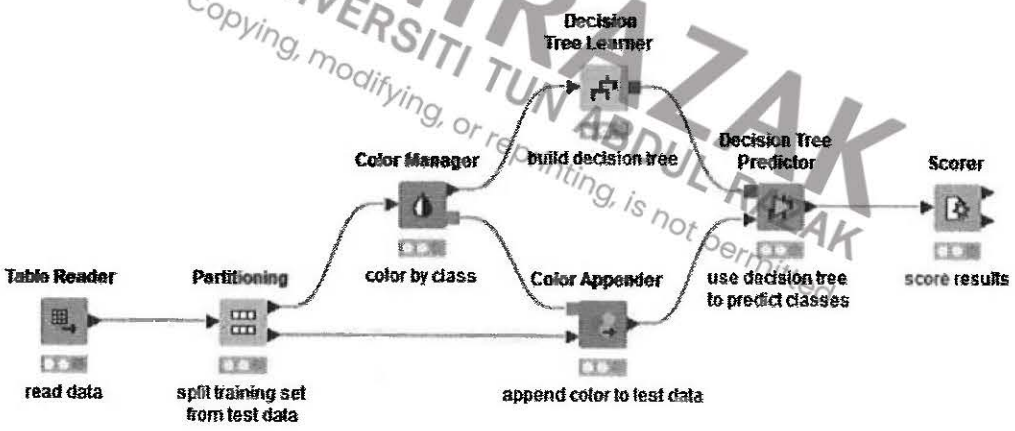
A. Because the algorithms have different objective functions. The logistic regression model optimizes itself to have a maximum margin, while the support vector machine model optimizes itself to have the probability of positive examples as close to 1 and the probability of negative examples as close to 0.
B. Because the algorithms have different loss functions. The logistic regression model uses mean squared error loss while the support vector machine model uses hinge loss.
C. Because the algorithms have different objective functions. The logistic regression model optimizes itself to have the probability of positive examples as close to 1 and the probability of negative examples as close to 0, while the support vector machine model optimizes itself to have a maximum margin.
D. Because the algorithms have different loss functions. The logistic regression model uses multi-class cross-entropy loss while the support vector machine model uses hinge loss

17. Given a KNIME workflow screenshot for a decision tree classification. Based on that, which node should you edit to change the dataset?



A. The table reader node
B. The scorer node
C. The decision tree learner node
D. The decision tree predictor node

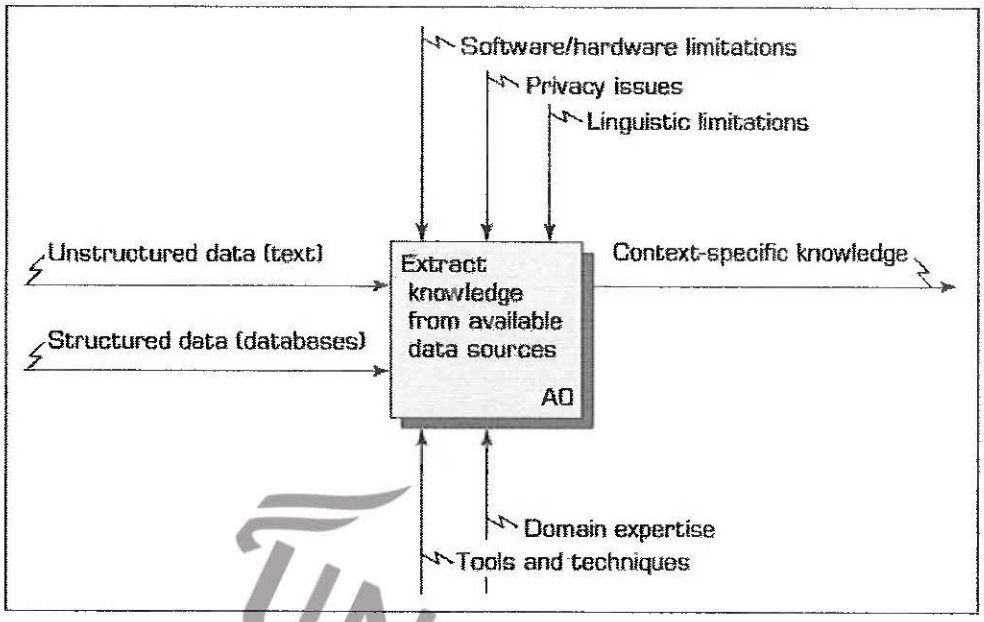
18. Given a KNIME workflow screenshot for a decision tree classification. Based on that, what are the two inputs of the Decision Tree Predictor node?



A. Untrained decision tree learner and colour-coded test data
B. Trained decision tree model and colour-coded train data
C. Trained decision tree model and colour-coded test data
D. Untrained decision tree learner and colour-coded train data

19. Assume that you have implemented a text mining project by following the process of the figure shown below.

    Which of the following information will you present to the decision-makers in the company to inform them about the insights on the latest product?



A. Structured
B. Context-specific knowledge
C. Unstructured
D. Linguistic limitations

20. Assume that you are hired by the Medical Research Centre A to implement data mining on its X-ray data. As the medical centre is having a shortage of medical staffs, you are asked to implement a machine learning solution to automatically screen all the X-rays and only forward the ones with suspected lung cancer to the medical team for further investigation. However, the research centre can only provide you with normal lung X-rays for model training and mentioned that the lungs with cancer will look very different physically compared to the normal ones. Which machine learning technique is best to be implemented for the given scenario?

    A. Association Pattern Mining
    B. Data Classification
    C. Outlier Analysis
    D. All of the above

SECTION B **(60 Marks)**

There are THREE (3) questions in this part. Answer ALL questions in the answer booklet.

Question 1 **(30 marks)**

A database has 4 transactions, shown below.

TID	Date	items_bought
T100	10/15/04	{K, A, D, B}
T200	10/15/04	{D, A, C, E, B}
T300	10/19/04	{C, A, B, E}
T400	10/22/04	{B, A, D}

Given itemset: 1-itemset: {D}, 2-itemset: {B}, 3-itemset: {C,D}, 4-itemset: {B,D}.

i. State Absolute Support & Relative Support for each itemset. (12 marks)

ii. Given minsup: 60% and minconf:80%, find frequent itemset association for C → D, B → D. (8 marks)

iii. Briefly explain the data mining method that has been used in this question. (10 marks)

Question 2 **(30 marks)**

For each of the following situations, write down the algorithm that would be an appropriate solution to the given data mining problem. You do need to justify supporting your answer.

i. A telecommunications company wants to predict which customers are likely to switch to a competitor. (6 marks)

ii. A social media company wants to analyze user posts to gauge public sentiment on various topics. (6 marks)

iii. A healthcare provider wants to predict which patients are at risk for certain diseases. (6 marks)

iv. A cybersecurity firm wants to identify unusual network traffic that may indicate a cyber attack. (6 marks)

v. A manufacturing company wants to predict equipment failures before they occur. (6 marks)

***** END OF QUESTION PAPER *****