UNIRAZAK
UNIVERSITI TUN ABDUL RAZAK

FINAL EXAMINATION

NOVEMBER 2023

COURSE TITLE **DATA MINING**

COURSE CODE **RBAN3213**

DATE/DAY **20 FEBRUARY 2024 / TUESDAY**

TIME/DURATION **09:00 AM - 11:00 AM / 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 **11** Printed Pages including front page)

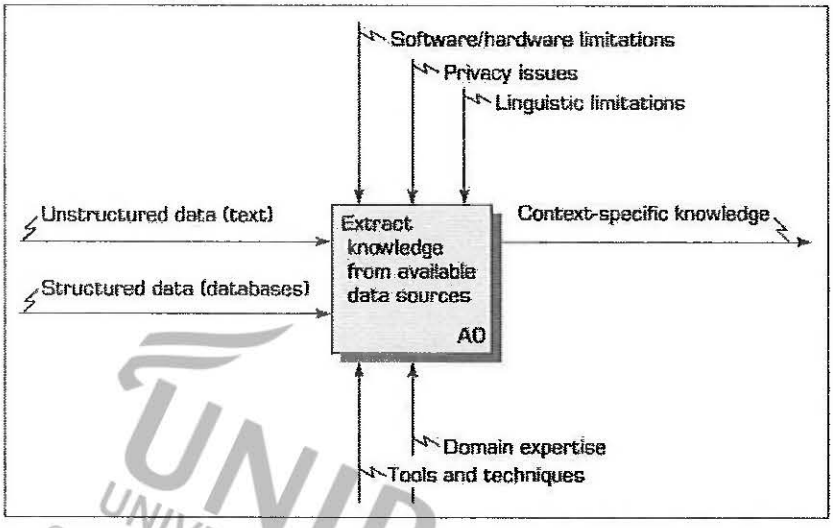
*****DO NOT OPEN THE QUESTION PAPER UNTIL YOU ARE TOLD TO DO SO*****

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.



Assume that you have implemented a text mining project by following the process of the figure shown above. 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 data
B. Unstructured data
C. Context-specific knowledge
D. Linguistic limitation

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

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

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

4. Assume that you are the project manager of the new social networking website project named TeamLike. The data scientist from your team has requested for an appropriate dataset to detect malicious login behaviors from users. You decided that the event logs are useful for the task as they contain a series of user activities at the Login page. What kind of data is that?

   A. Direct sequence and strings
   B. Spatiotemporal
   C. Network and graph
   D. Qualitative

5. Given the example:

   > The number of infected and recovered COVID-19 patients are recorded every day in every state of Malaysia to monitor the progress of the pandemic.

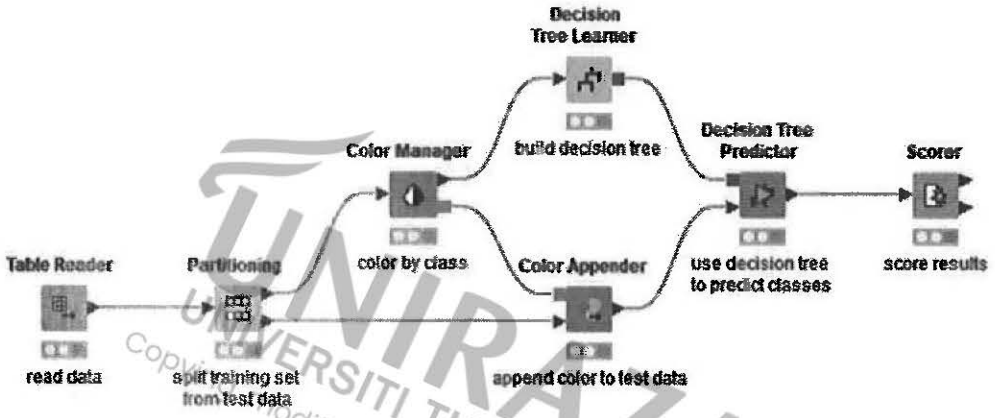
   Determine the contextual and behavioural attributes of the given example.

   A. Contextual: The number of hospitalized patients; Behavioural: The number of people who did not wear masks in the public
   B. Contextual: The number of people who did not wear masks in the public; Behavioural: The number of hospitalized patients
   C. Contextual: The number of infected and recovered patients; Behavioural: The date of incident and name of states
   D. Contextual: The date of incident and name of states; Behavioural: The number of infected and recovered patients

6. Assume that you are hired by the Medical Research Centre A to implement data mining on its X-ray data. As the medical center 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 esophagus cancer to the medical team for further investigation. However, the research center can only provide you with normal lung X-rays for model training and mentioned that the esophagus 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

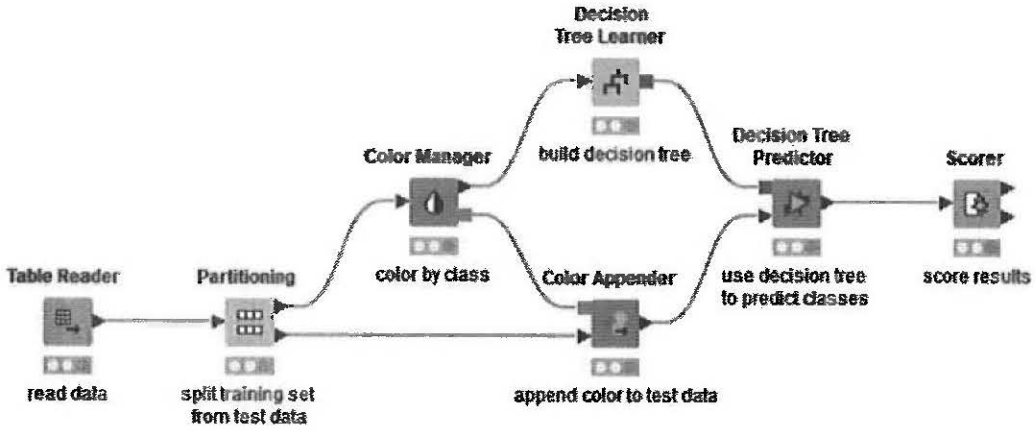
7.



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

8.



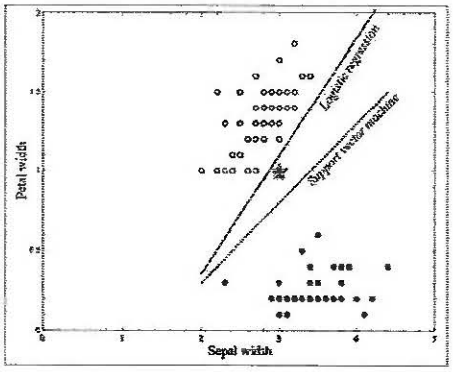
Given a KNIME workflow screenshot for a decision tree classification. Based on that, which node should you edit to change the dataset?

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

9. During a product meeting with all stakeholders, your colleague suggested enhancing the company's food delivery mobile application with a new feature, which is to suggest restaurants based on users' food ordering history. The project manager is concerned about the privacy and security issues of user accounts with the new feature implementation. Which aspect of the data mining implementation should the data team address to ensure a good coverage around the privacy and security topics of the user accounts?

A. Solution cost
B. Data governance
C. Processing capabilities
D. Data integration

10.



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 management asks: "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

11. The data team that you are managing has just received a new task, which is to implement natural language processing techniques in analyzing customer reviews on the latest product of the company. It turns out that the customer review data needs a lot more time to process than the team has anticipated. As a project manager presenting the progress during the monthly meeting of business stakeholders, which of the following is **NOT** part of your arguments about the challenges that your team is facing in processing the text data?

A. There are a mix of local and internet slang words in the reviews
B. The reviews are bad and the use of words are too intimidating
C. Some reviews are written in foreign languages, such as Japanese, that are hard to determine their word boundaries for processing
D. Some words have multiple meanings and they need additional steps to determine the correct meaning

12. Assume that you are working as a social media executive and you are considering enhancing your skill by taking a data mining course. How would having a data-analytics mindset help you in your career?

    A. I can easily exchange information with my data science colleagues by understanding how data is processed and analyzed
    B. I can take business problems in the social media department and figure out how to solve them with the help of structured data analysis
    C. When I evaluate a data report, I can tell if the methodologies, process and evaluation make sense
    D. All of the above

13. Your data team has just implemented a data model based on the CRISP-DM process. During the evaluation phase, you found out that the model does not solve the business problem. The business problem is that Boeing 747 Aircraft model tend to break down unexpectedly and that causes unfulfillment of the planned flight schedules. However, the task that was received by the data team was aimed at preventing the drop of sales of airline ticket purchase. At which phase of the CRISP-DM did the misunderstanding occur?

    A. Business understanding
    B. Data understanding
    C. Data preparation
    D. Modeling

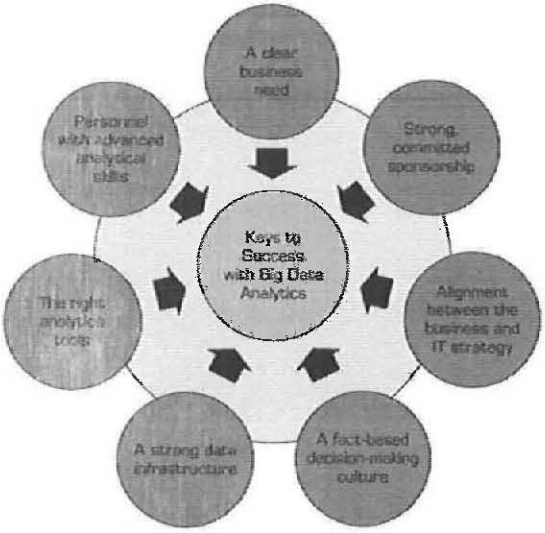
14. Kaely has a grocery store and she hired you to perform some data mining techniques to boost the sales outcome of her store. She told you that she would like to put frequently bought items close to each other on the shelves so that customers are more likely to purchase extra items. You implemented an association pattern mining model and found out there are too many frequent item sets. What would you do to decrease the number of frequent item sets most suitably?

    A. Observe customer purchase in the store and manually determine frequent item sets based on what you observed that day
    B. Increase the minimum support value of the association pattern mining model
    C. Decrease the minimum support value of the association pattern mining model
    D. Randomly select a few frequent item sets and ignore the others

15. Unlike a logistic regression model, a classification tree model uses divide-and-conquer approach and processes one feature at a time in creating its decision boundaries.

    A. True
    B. False

16. Answer the question below based on the diagram.



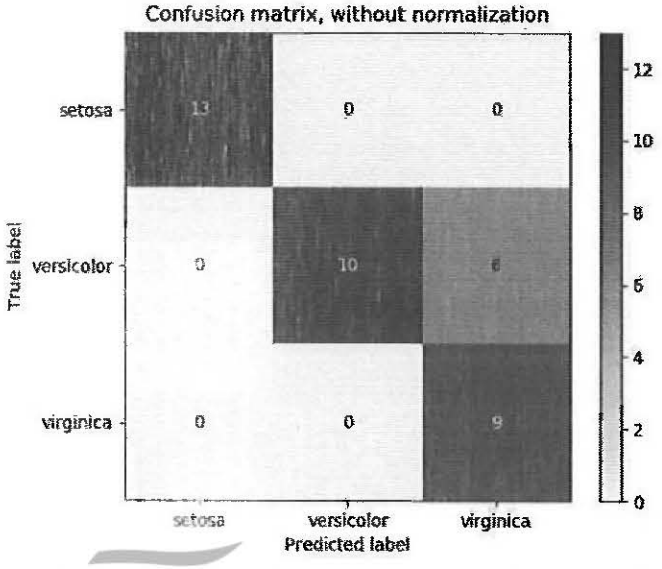
The main driver for Big Data Analytics should be_____.

A. A fact-based decision-making culture
B. A clear business need
C. The right analytics tools
D. A strong data infrastructure

17. 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

18.



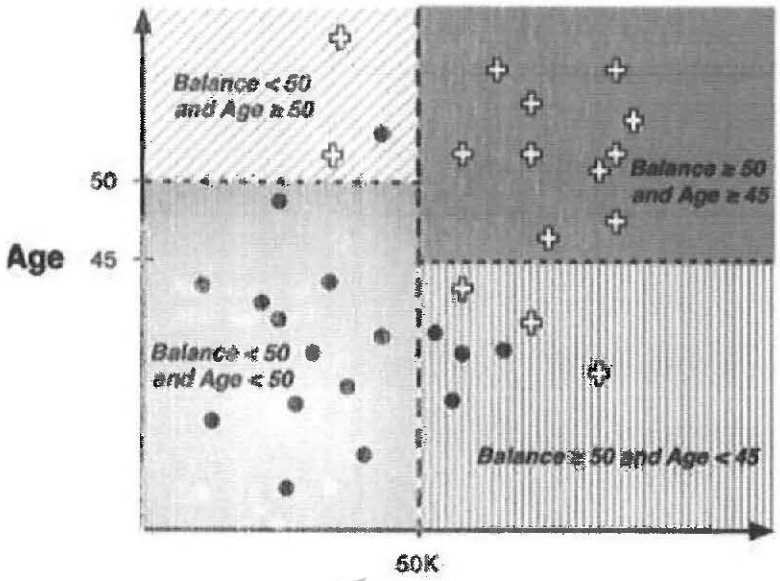
Assume that you have trained an Iris flower classification model with categories Setosa, Versicolor and Virginica. The outcome is plotted as a confusion matrix heatmap as shown above. Based on the heatmap, which category has the worst prediction outcome and what makes you conclude that?

A. Virginica, because 6 samples are wrongly predicted as Versicolor
B. Versicolor, because 6 samples are wrongly predicted as Virginica
C. Setosa, because 6 samples are wrongly predicted as Versicolor
D. Versicolor, because 6 samples are wrongly predicted as Setosa

19. Which algorithm is used to find correlations among different attributes in a data set?

A. Associative algorithm
B. Association algorithm
C. Time series algorithm
D. Series algorithm

20.



A classification tree as shown in the above diagram uses the divide and conquer approach, cutting up the instance space arbitrarily finely into very small regions, the decision boundaries.

A. True
B. False

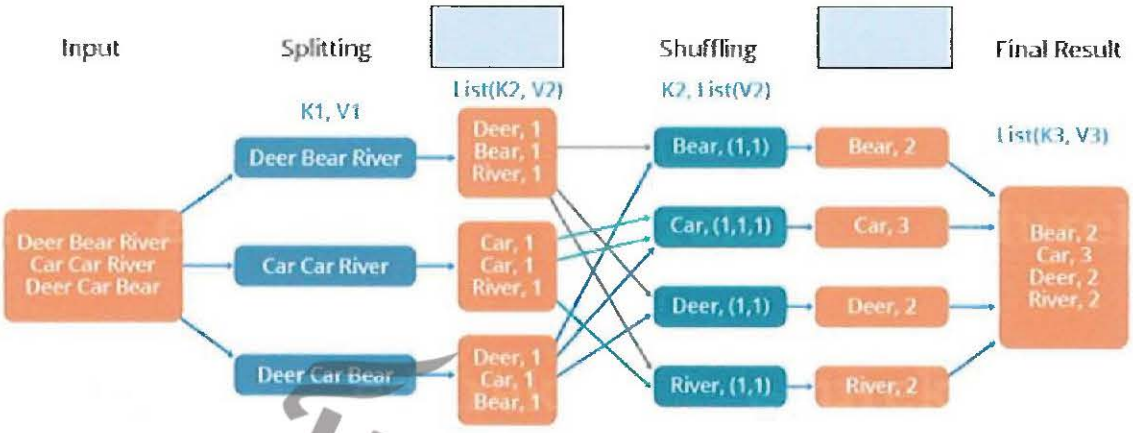
SECTION B **(60 Marks)**

There is One (1) question in this section. Answer ALL questions in an essay format.

The diagram shows an example model of MapReduce, a Big Data technology. Let's assume that we have a text file which has following content to perform a word count:

Dear, Bear, River, Car, Car, River, Deer, Car and Bear

The overall MapReduce process will look like:



a. Name the **TWO (2)** missing processes, step 3 and step 5 in the MapReduce process shown in the diagram above. (10 marks)

b. Describe each of the process and steps. You may start with how the input of the words has been split up until the final result step being written in the key, value format to the output file.

The steps should include all the 6 steps from the above diagram. (50 marks)

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