Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. What does the leaf node in decision tree indicates

A : sub tree

B : class label

C : testing node

D : condition

Q.no 2. sensitivity is also known as

A : false rate

B : recall

C : negative rate

D : recognition rate

Q.no 3. the negative tuples that were correctly labeled by the classifier

A : False positives(FP)

B : True positives(TP)

C : True negatives (TN)

D : False negatives(FN)

Q.no 4. Removing duplicate records is a process called

A : recovery

B : data cleaning

C : data cleansing

D : data pruning

Q.no 5. For Apriori algorithm, what is the first phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 6. Multi-class classification makes the assumption that each sample is assigned to

A : one and only one label

B : many labels

C : one or many labels

D : no label

Q.no 7. Multilevel association rules can be mined efficiently using

A : Support

B : Confidence

C : Support count

D : Concept Hierarchies under support-confidence framework

Q.no 8. What is the method to interpret the results after rule generation?

A : Absolute Mean

B : Lift ratio

C : Gini Index

D : Apriori

Q.no 9. Self-training is the simplest form of

A : supervised classification

B : semi-supervised classification

C : unsupervised classification

D : regression

Q.no 10. Which of the following is direct application of frequent itemset mining?

A : Social Network Analysis

B : Market Basket Analysis

C : Outlier Detection

D : Intrusion Detection

Q.no 11. Hidden knowledge referred to

A : A set of databases from different vendors, possibly using different database paradigms

B : An approach to a problem that is not guaranteed to work but performs well in most cases

C : Information that is hidden in a database and that cannot be recovered by a simple SQL query

D : None of these

Q.no 12. The schema is collection of stars. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 13. The Synonym for data mining is

A : Data warehouse

B : Knowledge discovery in database

C : ETL

D : Business Intelligemce

Q.no 14. Which of the following are methods for supervised classification?

A : Decision tree

B : K-Means

C : Hierarchical

D : Apriori

Q.no 15. These are the intermediate servers that stand in between a relational back-end server and client front-end tools

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 16. Color is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : numeric

Q.no 17. What are two steps of tree pruning work?

A : Pessimistic pruning and Optimistic pruning

B : Postpruning and Prepruning

C : Cost complexity pruning and time complexity pruning

D : None of the options

Q.no 18. A data cube is defined by

A : Dimensions

B : Facts

C : Dimensions and Facts

D : Dimensions or Facts

Q.no 19. For Apriori algorithm, what is the second phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 20. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 21. Lazy learner classification approach is

A : learner waits until the last minute before constructing model to classify

B : a given training data constructs a model first and then uses it to classify

C : the network is constructed by human experts

D : None of the options

Q.no 22. Cross validation involves

A : testing the machine on all possible ways by substituting the original sample into training set

B : testing the machine on all possible ways by dividing the original sample into training and validation sets.

C : testing the machine with only validation sets

D : testing the machine on only testing datasets.

Q.no 23. The rule is considered as intersting if

A : They satisfy both minimum support and minimum confidence threshold

B : They satisfy both maximum support and maximum confidence threshold

C : They satisfy maximum support and minimum confidence threshold

D : They satisfy minimum support and maximum confidence threshold

Q.no 24. Data independence means

A : Data is defined separately and not included in programs

B : Programs are not dependent on the physical attributes of the data

C : Programs are not dependent on the logiical attributes of the data

D : Programs are not dependent on the physical attributes as well as logical attributes of the data

Q.no 25. Which of the following is a predictive model?

A : Clustering

B : Regression

C : Summarization

D : Association rules

Q.no 26. The data cubes are generally

A : 1 Dimensional

B : 2 Dimensional

C : 3 Dimensional

D : n-Dimensional

Q.no 27. Identify the example of sequence data

A : weather forecast

B : data matrix

C : market basket data

D : genomic data

Q.no 28. The frequent-item-header-table consists of number fields

A : Only one

B : Two

C : Three

D : Four

Q.no 29. How are metarules useful in mining of association rules?

A : Allow users to specify threshold measures

B : Allow users to specify task relevant data

C : Allow users to specify the syntactic forms of rules

D : Allow users to specify correlation or association

Q.no 30. Which of the following activities is a data mining task?

A : Monitoring the heart rate of a patient for abnormalities

B : Extracting the frequencies of a sound wave

C : Predicting the outcomes of tossing a (fair) pair of dice

D : Dividing the customers of a company according to their profitability

Q.no 31. When do you consider an association rule interesting?

A : If it only satisfies minimum support

B : If it only satisfies minimum confidence

C : If it satisfies both minimum support and minimum confidence

D : There are other measures to check interesting rules

Q.no 32. What is the approach of basic algorithm for decision tree induction?

A : Greedy

B : Top Down

C : Procedural

D : Step by Step

Q.no 33. What do you mean by support(A)?

A : Total number of transactions containing A

B : Total Number of transactions not containing A

C : Number of transactions containing A / Total number of transactions

D : Number of transactions not containing A / Total number of transactions

Q.no 34. Which of the following probabilities are used in the Bayes theorem.

A : P(Ci|X)

B : P(Ci)

C : P(X|Ci)

D : P(X)

Q.no 35. In which step of Knowledge Discovery, multiple data sources are combined?

A : Data Cleaning

B : Data Integration

C : Data Selection

D : Data Transformation

Q.no 36. The Galaxy Schema is also called as

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 37. Handwritten digit recognition classifying an image of a handwritten number into a digit from 0 to 9 is example of

A : Multiclassification

B : Multi-label classification

C : Imbalanced classification

D : Binary Classification

Q.no 38. What type of data do you need for a chi-square test?

A : Categorical

B : Ordinal

C : Interval

D : Scales

Q.no 39. For a classification problem with highly imbalanced class. The majority class is observed 99% of times in the training data.
Your model has 99% accuracy after taking the predictions on test data. Which of the following is not true in such a case?

A : Imbalaced problems should not be measured using Accuracy metric.

B : Accuracy metric is not a good idea for imbalanced class problems.

C : Precision and recall metrics aren't good for imbalanced class problems.

D : Precision and recall metrics are good for imbalanced class problems.

Q.no 40. Which of the following property typically does not hold for similarity measures between two objects ?

A : Symmetry

B : Definiteness

C : Triangle inequality

D : Transitive

Q.no 41. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 42. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 43. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 44. Ordinal attribute has three distinct values such as Fair, Good, and Excellent.
If x and y are two objects of ordinal attribute with Fair and Good values respectively, then what is the distance from object y to x?

A : 1

B : 0

C : 0.5

D : 0.75

Q.no 45. Which of the following operation is requird to calculate cosine similarity?

A : Vector dot product

B : Exponent

C : Modulus

D : Percentage

Q.no 46. Which is the most well known association rule algorithm and is used in most commercial products.

A : Apriori algorithm

B : Pincer-search algorithm

C : Distributed algorithm

D : Partition algorithm

Q.no 47. What is the another name of Supremum distance?

A : Wighted Euclidean distance

B : City Block
distance

C : Chebyshev distance

D : Euclidean distance

Q.no 48. a model predicts 50 examples belonging to the minority class, 45 of which are true positives and five of which are false positives. Precision of model is

A : Precision= 0.90

B : Precision= 0.79

C : Precision= 0.45

D : Precision= 0.68

Q.no 49. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Q.no 50. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 51. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 52. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 53. Which operation data warehouse requires ?

A : Initial loading of data

B : Transaction processing

C : Recovery

D : Concurrency control mechanisms

Q.no 54. The problem of finding hidden structure from unlabeled data is called as

A : Supervised learning

B : Unsupervised learning

C : Reinforcement Learning

D : Semisupervised learning

Q.no 55. A model makes predictions and predicts 120 examples as belonging to the minority class, 90 of which are correct, and 30 of which are incorrect. Precision of model is

A : Precision = 0.89

B : Precision = 0.23

C : Precision = 0.45

D : Precision = 0.75

Q.no 56. Accuracy is

A : Number of correct predictions out of total no. of predictions

B : Number of incorrect predictions out of total no. of predictions

C : Number of predictions out of total no. of predictions

D : Total number of predictions

Q.no 57. What does a Pearson's product-moment allow you to identify?

A : Whether there is a relationship between variables

B : Whether there is a significant effect and interaction of independent variables

C : Whether there is a significant difference between variables

D : Whether there is a significant effect and interaction of dependent variables

Q.no 58. A model makes predictions and predicts 90 of the positive class predictions correctly and 10 incorrectly.Recall of model is

A : Recall=0.9

B : Recall=0.39

C : Recall=0.65

D : Recall=5.0

Q.no 59. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 60. These server performs the faster computation

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Answer for Question No 1. is b

Answer for Question No 2. is b

Answer for Question No 3. is c

Answer for Question No 4. is b

Answer for Question No 5. is c

Answer for Question No 6. is a

Answer for Question No 7. is d

Answer for Question No 8. is b

Answer for Question No 9. is b

Answer for Question No 10. is b

Answer for Question No 11. is c

Answer for Question No 12. is c

Answer for Question No 13. is b

Answer for Question No 14. is a

Answer for Question No 15. is a

Answer for Question No 16. is a

Answer for Question No 17. is b

Answer for Question No 18. is c

Answer for Question No 19. is a

Answer for Question No 20. is a

Answer for Question No 21. is a

Answer for Question No 22. is c

Answer for Question No 23. is a

Answer for Question No 24. is d

Answer for Question No 25. is b

Answer for Question No 26. is d

Answer for Question No 27. is d

Answer for Question No 28. is b

Answer for Question No 29. is c

Answer for Question No 30. is a

Answer for Question No 31. is c

Answer for Question No 32. is a

Answer for Question No 33. is c

Answer for Question No 34. is a

Answer for Question No 35. is b

Answer for Question No 36. is c

Answer for Question No 37. is a

Answer for Question No 38. is a

Answer for Question No 39. is c

Answer for Question No 40. is c

Answer for Question No 41. is a

Answer for Question No 42. is d

Answer for Question No 43. is b

Answer for Question No 44. is c

Answer for Question No 45. is a

Answer for Question No 46. is a

Answer for Question No 47. is c

Answer for Question No 48. is a

Answer for Question No 49. is b

Answer for Question No 50. is d

Answer for Question No 51. is d

Answer for Question No 52. is d

Answer for Question No 53. is a

Answer for Question No 54. is b

Answer for Question No 55. is d

Answer for Question No 56. is a

Answer for Question No 57. is a

Answer for Question No 58. is a

Answer for Question No 59. is a

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Postpruning is

A : Removing branches from fully grown tree

B : Stop constructing tree if this would result in the measure falling below a threshold

C : construting a new tree

D : Flow-Chart

Q.no 2. If two documents are similar, then what is the measure of angle between two documents?

A : 30

B : 60

C : 90

D : 0

Q.no 3. CART stands for

A : Regression

B : Classification

C : Classification and Regression Trees

D : Decision Trees

Q.no 4. The first steps involved in the knowledge discovery is?

A : Data Integration

B : Data Selection

C : Data Transformation

D : Data Cleaning

Q.no 5. These are the intermediate servers that stand in between a relational back-end server and client front-end tools

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 6. sensitivity is also known as

A : false rate

B : recall

C : negative rate

D : recognition rate

Q.no 7. Which of the following is not a type of constraints?

A : Data constraints

B : Rule constraints

C : Knowledge type constraints

D : Time constraints

Q.no 8. Baysian classification in based on

A : probability for the hypothesis

B : Support

C : tree induction

D : Trees

Q.no 9. Which one of the following is true for decision tree

A : Decision tree is useful in decision making

B : Decision tree is similar to OLTP

C : Decision Tree is similar to cluster analysis

D : Decision tree needs to find probabilities of hypothesis

Q.no 10. Hidden knowledge referred to

A : A set of databases from different vendors, possibly using different database paradigms

B : An approach to a problem that is not guaranteed to work but performs well in most cases

C : Information that is hidden in a database and that cannot be recovered by a simple SQL query

D : None of these

Q.no 11. What is an alternative form of Euclidean distance?

A : L1 norm

B : L2 norm

C : Lmax norm

D : L norm

Q.no 12. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 13. What are closed frequent itemsets?

A : A closed itemset

B : A frequent itemset

C : An itemset which is both closed and frequent

D : Not frequent itemset

Q.no 14. A decision tree is also known as

A : general tree

B : binary tree

C : prediction tree

D : None of the options

Q.no 15. cross-validation and bootstrap methods are common techniques for assessing

A : accuracy

B : Precision

C : recall

D : performance

Q.no 16. A multidimensional data model is typically organized around a central theme which is represented by

A : Dimension table

B : Fact table

C : Dimension table and Fact table

D : Dimension table or Fact table

Q.no 17. The problem of agents to learn from the environment by their interactions with dynamic environment is done in

A : Reinforcement learning

B : Multi-label classification

C : Binary Classification

D : Multiclassification

Q.no 18. Entropy is a measure of

A : impurity of an attribute

B : Purity of an attribute

C : Weight of an attribute

D : Class of an attribute

Q.no 19. the negative tuples that were correctly labeled by the classifier

A : False positives(FP)

B : True positives(TP)

C : True negatives (TN)

D : False negatives(FN)

Q.no 20. An ROC curve for a given model shows the trade-off between

A : random sampling

B : test data and train data

C : cross validation

D : the true positive rate (TPR) and the false positive rate (FPR)

Q.no 21. What is another name of data matrix?

A : Single mode

B : Two mode

C : Multi mode

D : Large mode

Q.no 22. Which of the following is a predictive model?

A : Clustering

B : Regression

C : Summarization

D : Association rules

Q.no 23. The rule is considered as intersting if

A : They satisfy both minimum support and minimum confidence threshold

B : They satisfy both maximum support and maximum confidence threshold

C : They satisfy maximum support and minimum confidence threshold

D : They satisfy minimum support and maximum confidence threshold

Q.no 24. Data independence means

A : Data is defined separately and not included in programs

B : Programs are not dependent on the physical attributes of the data

C : Programs are not dependent on the logiical attributes of the data

D : Programs are not dependent on the physical attributes as well as logical attributes of the data

Q.no 25. What do you mean by support(A)?

A : Total number of transactions containing A

B : Total Number of transactions not containing A

C : Number of transactions containing A / Total number of transactions

D : Number of transactions not containing A / Total number of transactions

Q.no 26. If first object X and Y coordinates are 3 and 5 respectively and second object X and Y coordinates are 10 and 3 respectively, then what is Manhattan disstance between these two objects?

A : 8

B : 13

C : 9

D : 10

Q.no 27. Number of records are comparatively more in

A : OLAP

B : OLTP

C : Same in OLAP and OLTP

D : Can not compare

Q.no 28. Which of the following operations are used to calculate proximity measures for ordinal attribute?

A : Replacement and discretization

B : Replacement and characterizarion

C : Replacement and normalization

D : Normalization and discretization

Q.no 29. Which of the following is necessary operation to calculate dissimilarity between ordinal attributes?

A : Replacement of ordinal categories

B : Correlation coefficient

C : Discretization

D : Randomization

Q.no 30. Multilevel association rule mining is

A : Association rules generated from candidate-generation method

B : Association rules generated from without candidate-generation method

C : Association rules generated from mining data at multiple abstarction level

D : Assocation rules generated from frequent itemsets

Q.no 31. In a decision tree each leaf node represents

A : Test conditions

B : Class labels

C : Attribute values

D : Decision

Q.no 32. The Galaxy Schema is also called as

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 33. For mining frequent itemsets, the Data format used by Apriori and FP-Growth algorithms are

A : Apriori uses horizontal and FP-Growth uses vertical data format

B : Apriori uses vertical and FP-Growth uses horizontal data format

C : Apriori and FP-Growth both uses vertical data format

D : Apriori and FP-Growth both uses horizontal data format

Q.no 34. The property of Apriori algorithm is

A : All nonempty subsets of a frequent itemsets must also be frequent

B : All empty subsets of a frequent itemsets must also be frequent

C : All nonempty subsets of a frequent itemsets must be not frequent

D : All nonempty subsets of a frequent itemsets can frequent or not frequent

Q.no 35. It is the main technique employed for data selection.

A : Noise

B : Sampling

C : Clustering

D : Histogram

Q.no 36. The probability of a hypothesis before the presentation of evidence is called as

A : Apriori probability

B : subjective probability

C : posterior probability

D : conditional probability

Q.no 37. In which step of Knowledge Discovery, multiple data sources are combined?

A : Data Cleaning

B : Data Integration

C : Data Selection

D : Data Transformation

Q.no 38. Some company wants to divide their customers into distinct groups to send offers this is an example of

A : Data Extraction

B : Data Classification

C : Data Discrimination

D : Data Selection

Q.no 39. The accuracy of a classifier on a given test set is the percentage of

A : test set tuples that are correctly classified by the classifier

B : test set tuples that are incorrectly classified by the classifier

C : test set tuples that are incorrectly misclassified by the classifier

D : test set tuples that are not classified by the classifier

Q.no 40. Which of the following is measure of document similarity?

A : Cosine dissimilarity

B : Sine similarity

C : Sine dissimilarity

D : Cosine similarity

Q.no 41. Which one of these is a tree based learner?

A : Rule based

B : Bayesian Belief Network

C : Bayesian classifier

D : Random Forest

Q.no 42. The problem of finding hidden structure from unlabeled data is called as

A : Supervised learning

B : Unsupervised learning

C : Reinforcement Learning

D : Semisupervised learning

Q.no 43. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 44. What is the range of the angle between two term frequency vectors?

A : Zero to Thirty

B : Zero to Ninety

C : Zero to One Eighty

D : Zero to Fourty Five

Q.no 45. Name the property of objects for which distance from first object to second and vice-versa is same.

A : Symmetry

B : Transitive

C : Positive definiteness

D : Traingle inequality

Q.no 46. Ordinal attribute has three distinct values such as Fair, Good, and Excellent.
If x and y are two objects of ordinal attribute with Fair and Good values respectively, then what is the distance from object y to x?

A : 1

B : 0

C : 0.5

D : 0.75

Q.no 47. A concept hierarchy that is a total or partial order among attributes in a database schema is called

A : Mixed hierarchy

B : Total hierarchy

C : Schema hierarchy

D : Concept generalization

Q.no 48. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 49. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Q.no 50. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 51. a model predicts 50 examples belonging to the minority class, 45 of which are true positives and five of which are false positives. Precision of model is

A : Precision= 0.90

B : Precision= 0.79

C : Precision= 0.45

D : Precision= 0.68

Q.no 52. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 53. High entropy means that the partitions in classification are

A : pure

B : Not pure

C : Useful

D : Not useful

Q.no 54. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 55. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 56. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 57. Correlation analysis is used for

A : handling missing values

B : identifying redundant attributes

C : handling different data formats

D : eliminating noise

Q.no 58. A data normalization technique for real-valued attributes that divides each numerical value by the same power of 10.

A : min-max normalization

B : z-score normalization

C : decimal scaling

D : decimal smoothing

Q.no 59. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 60. Holdout method, Cross-validation and Bootstrap methods are techniques to estimate

A : Precision

B : Classifier performance

C : Recall

D : F-measure

Answer for Question No 1. is a

Answer for Question No 2. is d

Answer for Question No 3. is c

Answer for Question No 4. is d

Answer for Question No 5. is a

Answer for Question No 6. is b

Answer for Question No 7. is d

Answer for Question No 8. is a

Answer for Question No 9. is a

Answer for Question No 10. is c

Answer for Question No 11. is b

Answer for Question No 12. is b

Answer for Question No 13. is c

Answer for Question No 14. is c

Answer for Question No 15. is a

Answer for Question No 16. is b

Answer for Question No 17. is a

Answer for Question No 18. is a

Answer for Question No 19. is c

Answer for Question No 20. is d

Answer for Question No 21. is b

Answer for Question No 22. is b

Answer for Question No 23. is a

Answer for Question No 24. is d

Answer for Question No 25. is c

Answer for Question No 26. is c

Answer for Question No 27. is b

Answer for Question No 28. is c

Answer for Question No 29. is a

Answer for Question No 30. is c

Answer for Question No 31. is b

Answer for Question No 32. is c

Answer for Question No 33. is d

Answer for Question No 34. is a

Answer for Question No 35. is b

Answer for Question No 36. is a

Answer for Question No 37. is b

Answer for Question No 38. is b

Answer for Question No 39. is a

Answer for Question No 40. is d

Answer for Question No 41. is d

Answer for Question No 42. is b

Answer for Question No 43. is a

Answer for Question No 44. is b

Answer for Question No 45. is a

Answer for Question No 46. is c

Answer for Question No 47. is c

Answer for Question No 48. is a

Answer for Question No 49. is b

Answer for Question No 50. is a

Answer for Question No 51. is a

Answer for Question No 52. is d

Answer for Question No 53. is b

Answer for Question No 54. is d

Answer for Question No 55. is b

Answer for Question No 56. is d

Answer for Question No 57. is b

Answer for Question No 58. is c

Answer for Question No 59. is a

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Which angle is used to measure document similarity?

A : Sin

B : Tan

C : Cos

D : Sec

Q.no 2. The first steps involved in the knowledge discovery is?

A : Data Integration

B : Data Selection

C : Data Transformation

D : Data Cleaning

Q.no 3. cross-validation and bootstrap methods are common techniques for assessing

A : accuracy

B : Precision

C : recall

D : performance

Q.no 4. The task of building decision model from labeled training data is called as

A : Supervised Learning

B : Unsupervised Learning

C : Reinforcement Learning

D : Structure Learning

Q.no 5. A multidimensional data model is typically organized around a central theme which is represented by

A : Dimension table

B : Fact table

C : Dimension table and Fact table

D : Dimension table or Fact table

Q.no 6. How can one represent document to calculate cosine similarity?

A : Vector

B : Matirx

C : List

D : Term frequency vector

Q.no 7. What is association rule mining?

A : Using association to find correlation rules

B : Same as frequent itemset mining

C : Finding of strong association rules using frequent itemsets

D : Finding of frequent itemset from large database

Q.no 8. What do you mean by dissimilarity measure of two objects?

A : Is a numerical measure of how alike two data objects are.

B : Is a numerical measure of how different two data objects are.

C : Higher when objects are more alike

D : Lower when objects are more different

Q.no 9. CART stands for

A : Regression

B : Classification

C : Classification and Regression Trees

D : Decision Trees

Q.no 10. OLAP database design is

A : Application-oriented

B : Object-oriented

C : Goal-oriented

D : Subject-oriented

Q.no 11. What is the method to interpret the results after rule generation?

A : Absolute Mean

B : Lift ratio

C : Gini Index

D : Apriori

Q.no 12. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 13. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 14. Color is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : numeric

Q.no 15. The schema is collection of stars. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 16. Data used to build a data mining model.

A : Validation Data

B : Training Data

C : Testing Data

D : Hidden Data

Q.no 17. The problem of agents to learn from the environment by their interactions with dynamic environment is done in

A : Reinforcement learning

B : Multi-label classification

C : Binary Classification

D : Multiclassification

Q.no 18. accuracy is used to measure

A : classifier’s true abilities

B : classifier’s analytic abilities

C : classifier’s decision abilities

D : classifier’s predictive abilities

Q.no 19. recall is a measure of

A : completeness of what percentage
of positive tuples are labeled

B : a measure of exactness for misclassification

C : a measure of exactness of what percentage of tuples are not classified

D : a measure of exactness of what percentage of tuples labeled as
negative are at actual

Q.no 20. Learning algorithm which trains with combination of labeled and unlabeled data.

A : Supervised

B : Unsupervised

C : Semi supervised

D : Non- supervised

Q.no 21. What is uniform support in multilevel association rule minig?

A : Use of minimum support

B : Use of minimum support and confidence

C : Use of same minimum threshold at each abstraction level

D : Use of minimum support and support count

Q.no 22. Which of the following activities is a data mining task?

A : Monitoring the heart rate of a patient for abnormalities

B : Extracting the frequencies of a sound wave

C : Predicting the outcomes of tossing a (fair) pair of dice

D : Dividing the customers of a company according to their profitability

Q.no 23. Which of the following operation is correct about supremum distance?

A : It gives maximum difference between any attribute of the objects

B : It gives minimum difference between any attribute of the objects

C : It gives maximum difference between fisrt attribute of the objects

D : It gives minimum difference between fisrt attribute of the objects

Q.no 24. Frequent patterns generated from association can be used for classification is called

A : Naïve Bays

B : Associative Classification

C : Preditctive Mining

D : Decision Tree

Q.no 25. Holdout and random subsampling are common techniques for assessing

A : K-Fold validation

B : cross validation

C : accuracy

D : sampling

Q.no 26. Which statement is true about the decision tree attribute selection process

A : A categorical attribute may appear in a tree node several times but a numeric attribute may appear at most once.

B : A numeric attribute may appear in several tree nodes but a categorical attribute may appear at most once.

C : Both numeric and categorical attributes may appear in several tree nodes.

D : Numeric and categorical attributes may appear in at most one tree node.

Q.no 27. Which of the following is not correct use of cross validation?

A : Selecting variables to include in a model

B : Comparing predictors

C : Selecting parameters in prediction function

D : classification

Q.no 28. In asymmetric attribute

A : No value is considered important over other values

B : All values are equal

C : Only non-zero value is important

D : Range of values is important

Q.no 29. When do you consider an association rule interesting?

A : If it only satisfies minimum support

B : If it only satisfies minimum confidence

C : If it satisfies both minimum support and minimum confidence

D : There are other measures to check interesting rules

Q.no 30. How will you counter over-fitting in decision tree?

A : By creating new rules

B : By pruning the longer rules

C : Both By pruning the longer rules' and ' By creating new rules'

D : BY creating new tree

Q.no 31. It is the main technique employed for data selection.

A : Noise

B : Sampling

C : Clustering

D : Histogram

Q.no 32. If A, B are two sets of items, and A is a subset of B. Which of the following statement is always true?

A : Support(A) is less than or equal to Support(B)

B : Support(A) is greater than or equal to Support(B)

C : Support(A) is equal to Support(B)

D : Support(A) is not equal to Support(B)

Q.no 33. Which is the wrong combination.

A : True negative=correctly indentified

B : False negative=incorrectly identified

C : False positive=correctly identified

D : True positive=correctly identified

Q.no 34. The data cubes are generally

A : 1 Dimensional

B : 2 Dimensional

C : 3 Dimensional

D : n-Dimensional

Q.no 35. A nearest neighbor approach is best used

A : with large-sized datasets.

B : when irrelevant attributes have been removed from the data.

C : when a generalized model of the data is desireable.

D : when an explanation of what has been found is of primary importance.

Q.no 36. The confusion matrix is a useful tool for analyzing

A : Regression

B : Classification

C : Sampling

D : Cross validation

Q.no 37. The rule is considered as intersting if

A : They satisfy both minimum support and minimum confidence threshold

B : They satisfy both maximum support and maximum confidence threshold

C : They satisfy maximum support and minimum confidence threshold

D : They satisfy minimum support and maximum confidence threshold

Q.no 38. What type of data do you need for a chi-square test?

A : Categorical

B : Ordinal

C : Interval

D : Scales

Q.no 39. Sensitivity is also referred to as

A : misclassification rate

B : true negative rate

C : True positive rate

D : correctness

Q.no 40. Number of records are comparatively more in

A : OLAP

B : OLTP

C : Same in OLAP and OLTP

D : Can not compare

Q.no 41. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Q.no 42. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Q.no 43. This technique uses mean and standard deviation scores to transform real-valued attributes.

A : decimal scaling

B : min-max normalization

C : z-score normalization

D : logarithmic normalization

Q.no 44. Consider three itemsets V1={tomato, potato,onion}, V2={tomato,potato}, V3={tomato}. Which of the following statement is correct?

A : support(V1) is greater than support (V2)

B : support(V3) is greater than support (V2)

C : support(V1) is greater than support(V3)

D : support(V2) is greater than support(V3)

Q.no 45. Which one of these is a tree based learner?

A : Rule based

B : Bayesian Belief Network

C : Bayesian classifier

D : Random Forest

Q.no 46. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 47. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 48. In Binning, we first sort data and partition into (equal-frequency) bins and then which of the following is not valid step

A : smooth by bin boundaries

B : smooth by bin median

C : smooth by bin means

D : smooth by bin values

Q.no 49. A model makes predictions and predicts 90 of the positive class predictions correctly and 10 incorrectly.Recall of model is

A : Recall=0.9

B : Recall=0.39

C : Recall=0.65

D : Recall=5.0

Q.no 50. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 51. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 52. Which is the most well known association rule algorithm and is used in most commercial products.

A : Apriori algorithm

B : Pincer-search algorithm

C : Distributed algorithm

D : Partition algorithm

Q.no 53. Name the property of objects for which distance from first object to second and vice-versa is same.

A : Symmetry

B : Transitive

C : Positive definiteness

D : Traingle inequality

Q.no 54. What does a Pearson's product-moment allow you to identify?

A : Whether there is a relationship between variables

B : Whether there is a significant effect and interaction of independent variables

C : Whether there is a significant difference between variables

D : Whether there is a significant effect and interaction of dependent variables

Q.no 55. These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

A : 25

B : 210

C : 62

D : 30

Q.no 56. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 57. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 58. Accuracy is

A : Number of correct predictions out of total no. of predictions

B : Number of incorrect predictions out of total no. of predictions

C : Number of predictions out of total no. of predictions

D : Total number of predictions

Q.no 59. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 60. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Answer for Question No 1. is c

Answer for Question No 2. is d

Answer for Question No 3. is a

Answer for Question No 4. is a

Answer for Question No 5. is b

Answer for Question No 6. is d

Answer for Question No 7. is c

Answer for Question No 8. is b

Answer for Question No 9. is c

Answer for Question No 10. is d

Answer for Question No 11. is b

Answer for Question No 12. is b

Answer for Question No 13. is a

Answer for Question No 14. is a

Answer for Question No 15. is c

Answer for Question No 16. is b

Answer for Question No 17. is a

Answer for Question No 18. is d

Answer for Question No 19. is a

Answer for Question No 20. is c

Answer for Question No 21. is c

Answer for Question No 22. is a

Answer for Question No 23. is a

Answer for Question No 24. is b

Answer for Question No 25. is c

Answer for Question No 26. is b

Answer for Question No 27. is d

Answer for Question No 28. is c

Answer for Question No 29. is c

Answer for Question No 30. is b

Answer for Question No 31. is b

Answer for Question No 32. is b

Answer for Question No 33. is c

Answer for Question No 34. is d

Answer for Question No 35. is b

Answer for Question No 36. is b

Answer for Question No 37. is a

Answer for Question No 38. is a

Answer for Question No 39. is c

Answer for Question No 40. is b

Answer for Question No 41. is b

Answer for Question No 42. is d

Answer for Question No 43. is c

Answer for Question No 44. is b

Answer for Question No 45. is d

Answer for Question No 46. is b

Answer for Question No 47. is c

Answer for Question No 48. is d

Answer for Question No 49. is a

Answer for Question No 50. is a

Answer for Question No 51. is d

Answer for Question No 52. is a

Answer for Question No 53. is a

Answer for Question No 54. is a

Answer for Question No 55. is d

Answer for Question No 56. is d

Answer for Question No 57. is b

Answer for Question No 58. is a

Answer for Question No 59. is d

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. How can one represent document to calculate cosine similarity?

A : Vector

B : Matirx

C : List

D : Term frequency vector

Q.no 2. In Data Characterization, class under study is called as?

A : Study Class

B : Intial Class

C : Target Class

D : Final Class

Q.no 3. What do you mean by dissimilarity measure of two objects?

A : Is a numerical measure of how alike two data objects are.

B : Is a numerical measure of how different two data objects are.

C : Higher when objects are more alike

D : Lower when objects are more different

Q.no 4. the negative tuples that were correctly labeled by the classifier

A : False positives(FP)

B : True positives(TP)

C : True negatives (TN)

D : False negatives(FN)

Q.no 5. A person trained to interact with a human expert in order to capture their knowledge.

A : knowledge programmer

B : knowledge developer

C : knowledge engineer

D : knowledge extractor

Q.no 6. Removing duplicate records is a process called

A : recovery

B : data cleaning

C : data cleansing

D : data pruning

Q.no 7. Self-training is the simplest form of

A : supervised classification

B : semi-supervised classification

C : unsupervised classification

D : regression

Q.no 8. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 9. recall is a measure of

A : completeness of what percentage
of positive tuples are labeled

B : a measure of exactness for misclassification

C : a measure of exactness of what percentage of tuples are not classified

D : a measure of exactness of what percentage of tuples labeled as
negative are at actual

Q.no 10. The task of building decision model from labeled training data is called as

A : Supervised Learning

B : Unsupervised Learning

C : Reinforcement Learning

D : Structure Learning

Q.no 11. The first steps involved in the knowledge discovery is?

A : Data Integration

B : Data Selection

C : Data Transformation

D : Data Cleaning

Q.no 12. sensitivity is also known as

A : false rate

B : recall

C : negative rate

D : recognition rate

Q.no 13. A decision tree is also known as

A : general tree

B : binary tree

C : prediction tree

D : None of the options

Q.no 14. Supervised learning and unsupervised clustering both require at least one

A : hidden attribute

B : output attribute

C : input attribute

D : categorical attribute

Q.no 15. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 16. Which angle is used to measure document similarity?

A : Sin

B : Tan

C : Cos

D : Sec

Q.no 17. Hidden knowledge referred to

A : A set of databases from different vendors, possibly using different database paradigms

B : An approach to a problem that is not guaranteed to work but performs well in most cases

C : Information that is hidden in a database and that cannot be recovered by a simple SQL query

D : None of these

Q.no 18. The example of knowledge type constraints in constraint based mining is

A : Association or Correlation

B : Rule templates

C : Task relevant data

D : Threshold measures

Q.no 19. Which technique finds the frequent itemsets in just two database scans?

A : Partitioning

B : Sampling

C : Hashing

D : Dynamic itemset counting

Q.no 20. A data matrix in which attributes are of the same type and asymmetric is called

A : Pattern matrix

B : Sparse data matrix

C : Document term matrix

D : Normal matrix

Q.no 21. Specificity is also referred to as

A : true negative rate

B : correctness

C : misclassification rate

D : True positive rate

Q.no 22. If first object X and Y coordinates are 3 and 5 respectively and second object X and Y coordinates are 10 and 3 respectively, then what is Manhattan disstance between these two objects?

A : 8

B : 13

C : 9

D : 10

Q.no 23. Which of the following property typically does not hold for similarity measures between two objects ?

A : Symmetry

B : Definiteness

C : Triangle inequality

D : Transitive

Q.no 24. The property of Apriori algorithm is

A : All nonempty subsets of a frequent itemsets must also be frequent

B : All empty subsets of a frequent itemsets must also be frequent

C : All nonempty subsets of a frequent itemsets must be not frequent

D : All nonempty subsets of a frequent itemsets can frequent or not frequent

Q.no 25. One of the most well known software used for classification is

A : Java

B : C4.5

C : Oracle

D : C++

Q.no 26. This supervised learning technique can process both numeric and categorical input attributes.

A : linear regression

B : Bayes classifier

C : logistic regression

D : backpropagation learning

Q.no 27. A lattice of cuboids is called as

A : Data cube

B : Dimesnion lattice

C : Master lattice

D : Fact table

Q.no 28. K-fold Cross Validation envisages

A : partitioning of the original sample into one sample.

B : partitioning of the original sample into ‘k’ equal sized sub-samples.

C : partitioning of the original sample into ‘k’ unequal sized sub-samples.

D : partitioning of the original sample into ‘k’ random samples.

Q.no 29. The fact table contains

A : The names of the facts

B : Keys to each of the related dimension tables

C : Facts and keys

D : Facts or keys

Q.no 30. In asymmetric attribute

A : No value is considered important over other values

B : All values are equal

C : Only non-zero value is important

D : Range of values is important

Q.no 31. Which of the following operation is correct about supremum distance?

A : It gives maximum difference between any attribute of the objects

B : It gives minimum difference between any attribute of the objects

C : It gives maximum difference between fisrt attribute of the objects

D : It gives minimum difference between fisrt attribute of the objects

Q.no 32. What type of matrix is required to represent binary data for proximity measures?

A : Normal matrix

B : Sparse matrix

C : Dense matrix

D : Contingency matrix

Q.no 33. Sensitivity is also referred to as

A : misclassification rate

B : true negative rate

C : True positive rate

D : correctness

Q.no 34. What is the limitation behind rule generation in Apriori algorithm?

A : Need to generate a huge number of candidate sets

B : Need to repeatedly scan the whole database and Check a large set of candidates by pattern matching

C : Dropping itemsets with valued information

D : Both (a) dnd (b)

Q.no 35. If A, B are two sets of items, and A is a subset of B. Which of the following statement is always true?

A : Support(A) is less than or equal to Support(B)

B : Support(A) is greater than or equal to Support(B)

C : Support(A) is equal to Support(B)

D : Support(A) is not equal to Support(B)

Q.no 36. Which of the following sequence is used to calculate proximity measures for ordinal attribute?

A : Replacement discretization and distance measure

B : Replacement characterizarion and distance measure

C : Normalization discretization and distance measure

D : Replacement normalization and distance measure

Q.no 37. For a classification problem with highly imbalanced class. The majority class is observed 99% of times in the training data.
Your model has 99% accuracy after taking the predictions on test data. Which of the following is not true in such a case?

A : Imbalaced problems should not be measured using Accuracy metric.

B : Accuracy metric is not a good idea for imbalanced class problems.

C : Precision and recall metrics aren't good for imbalanced class problems.

D : Precision and recall metrics are good for imbalanced class problems.

Q.no 38. Some company wants to divide their customers into distinct groups to send offers this is an example of

A : Data Extraction

B : Data Classification

C : Data Discrimination

D : Data Selection

Q.no 39. This operation may add new dimension to the cube

A : Roll up

B : Drill down

C : Slice

D : Dice

Q.no 40. What is another name of data matrix?

A : Single mode

B : Two mode

C : Multi mode

D : Large mode

Q.no 41. Holdout method, Cross-validation and Bootstrap methods are techniques to estimate

A : Precision

B : Classifier performance

C : Recall

D : F-measure

Q.no 42. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 43. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 44. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 45. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 46. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 47. This technique uses mean and standard deviation scores to transform real-valued attributes.

A : decimal scaling

B : min-max normalization

C : z-score normalization

D : logarithmic normalization

Q.no 48. The problem of finding hidden structure from unlabeled data is called as

A : Supervised learning

B : Unsupervised learning

C : Reinforcement Learning

D : Semisupervised learning

Q.no 49. These server performs the faster computation

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 50. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 51. High entropy means that the partitions in classification are

A : pure

B : Not pure

C : Useful

D : Not useful

Q.no 52. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 53. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 54. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 55. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 56. a model predicts 50 examples belonging to the minority class, 45 of which are true positives and five of which are false positives. Precision of model is

A : Precision= 0.90

B : Precision= 0.79

C : Precision= 0.45

D : Precision= 0.68

Q.no 57. Consider three itemsets V1={tomato, potato,onion}, V2={tomato,potato}, V3={tomato}. Which of the following statement is correct?

A : support(V1) is greater than support (V2)

B : support(V3) is greater than support (V2)

C : support(V1) is greater than support(V3)

D : support(V2) is greater than support(V3)

Q.no 58. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Q.no 59. A concept hierarchy that is a total or partial order among attributes in a database schema is called

A : Mixed hierarchy

B : Total hierarchy

C : Schema hierarchy

D : Concept generalization

Q.no 60. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Answer for Question No 1. is d

Answer for Question No 2. is c

Answer for Question No 3. is b

Answer for Question No 4. is c

Answer for Question No 5. is c

Answer for Question No 6. is b

Answer for Question No 7. is b

Answer for Question No 8. is a

Answer for Question No 9. is a

Answer for Question No 10. is a

Answer for Question No 11. is d

Answer for Question No 12. is b

Answer for Question No 13. is c

Answer for Question No 14. is c

Answer for Question No 15. is b

Answer for Question No 16. is c

Answer for Question No 17. is c

Answer for Question No 18. is a

Answer for Question No 19. is a

Answer for Question No 20. is b

Answer for Question No 21. is a

Answer for Question No 22. is c

Answer for Question No 23. is c

Answer for Question No 24. is a

Answer for Question No 25. is b

Answer for Question No 26. is b

Answer for Question No 27. is a

Answer for Question No 28. is d

Answer for Question No 29. is c

Answer for Question No 30. is c

Answer for Question No 31. is a

Answer for Question No 32. is d

Answer for Question No 33. is c

Answer for Question No 34. is d

Answer for Question No 35. is b

Answer for Question No 36. is d

Answer for Question No 37. is c

Answer for Question No 38. is b

Answer for Question No 39. is b

Answer for Question No 40. is b

Answer for Question No 41. is b

Answer for Question No 42. is a

Answer for Question No 43. is b

Answer for Question No 44. is a

Answer for Question No 45. is b

Answer for Question No 46. is d

Answer for Question No 47. is c

Answer for Question No 48. is b

Answer for Question No 49. is b

Answer for Question No 50. is a

Answer for Question No 51. is b

Answer for Question No 52. is a

Answer for Question No 53. is d

Answer for Question No 54. is d

Answer for Question No 55. is d

Answer for Question No 56. is a

Answer for Question No 57. is b

Answer for Question No 58. is d

Answer for Question No 59. is c

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. The problem of agents to learn from the environment by their interactions with dynamic environment is done in

A : Reinforcement learning

B : Multi-label classification

C : Binary Classification

D : Multiclassification

Q.no 2. Baysian classification in based on

A : probability for the hypothesis

B : Support

C : tree induction

D : Trees

Q.no 3. Which of the following is correct about Proximity measures?

A : Similarity

B : Dissimilarity

C : Similarity as well as Dissimilarity

D : Neither similarity nor dissimilarity

Q.no 4. For Apriori algorithm, what is the second phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 5. Learning algorithm which trains with combination of labeled and unlabeled data.

A : Supervised

B : Unsupervised

C : Semi supervised

D : Non- supervised

Q.no 6. The most widely used metrics and tools to assess a classification model are:

A : Conusion Matrix

B : Support

C : Entropy

D : Probability

Q.no 7. The schema is collection of stars. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 8. An ROC curve for a given model shows the trade-off between

A : random sampling

B : test data and train data

C : cross validation

D : the true positive rate (TPR) and the false positive rate (FPR)

Q.no 9. Multilevel association rules can be mined efficiently using

A : Support

B : Confidence

C : Support count

D : Concept Hierarchies under support-confidence framework

Q.no 10. Which of the following is not a type of constraints?

A : Data constraints

B : Rule constraints

C : Knowledge type constraints

D : Time constraints

Q.no 11. Data matrix is also called as

A : Object by object structure

B : Object by attribute structure

C : Attribute by attribute structure

D : Attribute by object structure

Q.no 12. Each dimension is represented by only one table. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 13. How can one represent document to calculate cosine similarity?

A : Vector

B : Matirx

C : List

D : Term frequency vector

Q.no 14. What is the method to interpret the results after rule generation?

A : Absolute Mean

B : Lift ratio

C : Gini Index

D : Apriori

Q.no 15. CART stands for

A : Regression

B : Classification

C : Classification and Regression Trees

D : Decision Trees

Q.no 16. sensitivity is also known as

A : false rate

B : recall

C : negative rate

D : recognition rate

Q.no 17. Height is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : Numeric

Q.no 18. cross-validation and bootstrap methods are common techniques for assessing

A : accuracy

B : Precision

C : recall

D : performance

Q.no 19. recall is a measure of

A : completeness of what percentage
of positive tuples are labeled

B : a measure of exactness for misclassification

C : a measure of exactness of what percentage of tuples are not classified

D : a measure of exactness of what percentage of tuples labeled as
negative are at actual

Q.no 20. OLAP database design is

A : Application-oriented

B : Object-oriented

C : Goal-oriented

D : Subject-oriented

Q.no 21. Every key structure in the data warehouse contains a time element

A : records

B : Explicitly

C : Implicitly and explicitly

D : Implicitly or explicitly

Q.no 22. This supervised learning technique can process both numeric and categorical input attributes.

A : linear regression

B : Bayes classifier

C : logistic regression

D : backpropagation learning

Q.no 23. For mining frequent itemsets, the Data format used by Apriori and FP-Growth algorithms are

A : Apriori uses horizontal and FP-Growth uses vertical data format

B : Apriori uses vertical and FP-Growth uses horizontal data format

C : Apriori and FP-Growth both uses vertical data format

D : Apriori and FP-Growth both uses horizontal data format

Q.no 24. How are metarules useful in mining of association rules?

A : Allow users to specify threshold measures

B : Allow users to specify task relevant data

C : Allow users to specify the syntactic forms of rules

D : Allow users to specify correlation or association

Q.no 25. A frequent pattern tree is a tree structure consisting of

A : A frequent-item-node

B : An item-prefix-tree

C : A frequent-item-header table

D : both B and C

Q.no 26. Learning with a complete system in mind with reference to interactions among
the systems and subsystems with proper understanding of systemic boundaries is

A : Multi-label classification

B : Reinforcement learning

C : Systemic learning

D : Machine Learning

Q.no 27. Handwritten digit recognition classifying an image of a handwritten number into a digit from 0 to 9 is example of

A : Multiclassification

B : Multi-label classification

C : Imbalanced classification

D : Binary Classification

Q.no 28. Which of the following activities is a data mining task?

A : Monitoring the heart rate of a patient for abnormalities

B : Extracting the frequencies of a sound wave

C : Predicting the outcomes of tossing a (fair) pair of dice

D : Dividing the customers of a company according to their profitability

Q.no 29. The frequent-item-header-table consists of number fields

A : Only one

B : Two

C : Three

D : Four

Q.no 30. The rule is considered as intersting if

A : They satisfy both minimum support and minimum confidence threshold

B : They satisfy both maximum support and maximum confidence threshold

C : They satisfy maximum support and minimum confidence threshold

D : They satisfy minimum support and maximum confidence threshold

Q.no 31. What is the limitation behind rule generation in Apriori algorithm?

A : Need to generate a huge number of candidate sets

B : Need to repeatedly scan the whole database and Check a large set of candidates by pattern matching

C : Dropping itemsets with valued information

D : Both (a) dnd (b)

Q.no 32. What is another name of data matrix?

A : Single mode

B : Two mode

C : Multi mode

D : Large mode

Q.no 33. How will you counter over-fitting in decision tree?

A : By creating new rules

B : By pruning the longer rules

C : Both By pruning the longer rules' and ' By creating new rules'

D : BY creating new tree

Q.no 34. In a decision tree each leaf node represents

A : Test conditions

B : Class labels

C : Attribute values

D : Decision

Q.no 35. A lattice of cuboids is called as

A : Data cube

B : Dimesnion lattice

C : Master lattice

D : Fact table

Q.no 36. If first object X and Y coordinates are 3 and 5 respectively and second object X and Y coordinates are 10 and 3 respectively, then what is Manhattan disstance between these two objects?

A : 8

B : 13

C : 9

D : 10

Q.no 37. one-versus-one(OVO) and one-versus-all (OVA) classification involves

A : more than two classes

B : Only two classes

C : Only one class

D : No class

Q.no 38. K-fold Cross Validation envisages

A : partitioning of the original sample into one sample.

B : partitioning of the original sample into ‘k’ equal sized sub-samples.

C : partitioning of the original sample into ‘k’ unequal sized sub-samples.

D : partitioning of the original sample into ‘k’ random samples.

Q.no 39. One of the most well known software used for classification is

A : Java

B : C4.5

C : Oracle

D : C++

Q.no 40. What do you mean by support(A)?

A : Total number of transactions containing A

B : Total Number of transactions not containing A

C : Number of transactions containing A / Total number of transactions

D : Number of transactions not containing A / Total number of transactions

Q.no 41. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 42. Accuracy is

A : Number of correct predictions out of total no. of predictions

B : Number of incorrect predictions out of total no. of predictions

C : Number of predictions out of total no. of predictions

D : Total number of predictions

Q.no 43. Which one of these is a tree based learner?

A : Rule based

B : Bayesian Belief Network

C : Bayesian classifier

D : Random Forest

Q.no 44. Which of the following operation is requird to calculate cosine similarity?

A : Vector dot product

B : Exponent

C : Modulus

D : Percentage

Q.no 45. Correlation analysis is used for

A : handling missing values

B : identifying redundant attributes

C : handling different data formats

D : eliminating noise

Q.no 46. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 47. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 48. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Q.no 49. What is the range of the angle between two term frequency vectors?

A : Zero to Thirty

B : Zero to Ninety

C : Zero to One Eighty

D : Zero to Fourty Five

Q.no 50. If True Positives (TP): 7, False Positives (FP): 1,False Negatives (FN): 4, True Negatives (TN): 18. Calculate Precision and Recall.

A : Precision = 0.88, Recall=0.64

B : Precision = 0.44, Recall=0.78

C : Precision = 0.88, Recall=0.22

D : Precision = 0.77, Recall=0.55

Q.no 51. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 52. In Binning, we first sort data and partition into (equal-frequency) bins and then which of the following is not valid step

A : smooth by bin boundaries

B : smooth by bin median

C : smooth by bin means

D : smooth by bin values

Q.no 53. A model makes predictions and predicts 90 of the positive class predictions correctly and 10 incorrectly.Recall of model is

A : Recall=0.9

B : Recall=0.39

C : Recall=0.65

D : Recall=5.0

Q.no 54. Name the property of objects for which distance from first object to second and vice-versa is same.

A : Symmetry

B : Transitive

C : Positive definiteness

D : Traingle inequality

Q.no 55. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 56. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 57. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Q.no 58. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 59. a model predicts 50 examples belonging to the minority class, 45 of which are true positives and five of which are false positives. Precision of model is

A : Precision= 0.90

B : Precision= 0.79

C : Precision= 0.45

D : Precision= 0.68

Q.no 60. Effectiveness of the browsing is highest. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Answer for Question No 1. is a

Answer for Question No 2. is a

Answer for Question No 3. is c

Answer for Question No 4. is a

Answer for Question No 5. is c

Answer for Question No 6. is a

Answer for Question No 7. is c

Answer for Question No 8. is d

Answer for Question No 9. is d

Answer for Question No 10. is d

Answer for Question No 11. is b

Answer for Question No 12. is a

Answer for Question No 13. is d

Answer for Question No 14. is b

Answer for Question No 15. is c

Answer for Question No 16. is b

Answer for Question No 17. is d

Answer for Question No 18. is a

Answer for Question No 19. is a

Answer for Question No 20. is d

Answer for Question No 21. is d

Answer for Question No 22. is b

Answer for Question No 23. is d

Answer for Question No 24. is c

Answer for Question No 25. is d

Answer for Question No 26. is c

Answer for Question No 27. is a

Answer for Question No 28. is a

Answer for Question No 29. is b

Answer for Question No 30. is a

Answer for Question No 31. is d

Answer for Question No 32. is b

Answer for Question No 33. is b

Answer for Question No 34. is b

Answer for Question No 35. is a

Answer for Question No 36. is c

Answer for Question No 37. is a

Answer for Question No 38. is d

Answer for Question No 39. is b

Answer for Question No 40. is c

Answer for Question No 41. is d

Answer for Question No 42. is a

Answer for Question No 43. is d

Answer for Question No 44. is a

Answer for Question No 45. is b

Answer for Question No 46. is a

Answer for Question No 47. is a

Answer for Question No 48. is b

Answer for Question No 49. is b

Answer for Question No 50. is a

Answer for Question No 51. is c

Answer for Question No 52. is d

Answer for Question No 53. is a

Answer for Question No 54. is a

Answer for Question No 55. is d

Answer for Question No 56. is b

Answer for Question No 57. is d

Answer for Question No 58. is b

Answer for Question No 59. is a

Answer for Question No 60. is a

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Which angle is used to measure document similarity?

A : Sin

B : Tan

C : Cos

D : Sec

Q.no 2. Data mining is best described as the process of

A : identifying patterns in data

B : deducing relationships in data

C : representing data

D : simulating trends in data

Q.no 3. These are the intermediate servers that stand in between a relational back-end server and client front-end tools

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 4. What does the leaf node in decision tree indicates

A : sub tree

B : class label

C : testing node

D : condition

Q.no 5. A multidimensional data model is typically organized around a central theme which is represented by

A : Dimension table

B : Fact table

C : Dimension table and Fact table

D : Dimension table or Fact table

Q.no 6. Data used to build a data mining model.

A : Validation Data

B : Training Data

C : Testing Data

D : Hidden Data

Q.no 7. The Synonym for data mining is

A : Data warehouse

B : Knowledge discovery in database

C : ETL

D : Business Intelligemce

Q.no 8. Color is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : numeric

Q.no 9. Cotraining is one form of

A : sampling

B : Reinforcement learning

C : unsupervised classification

D : semi-supervised classification

Q.no 10. What is C4.5 is used to build

A : Decision tree

B : Regression Analysis

C : Induction

D : Association Rules

Q.no 11. Training process that generates tree is called as

A : Pruning

B : Rule generation

C : Induction

D : spliiting

Q.no 12. Learning algorithm which trains with combination of labeled and unlabeled data.

A : Supervised

B : Unsupervised

C : Semi supervised

D : Non- supervised

Q.no 13. Which of the following is not frequent pattern?

A : Itemsets

B : Subsequences

C : Substructures

D : Associations

Q.no 14. What is an alternative form of Euclidean distance?

A : L1 norm

B : L2 norm

C : Lmax norm

D : L norm

Q.no 15. Which one of the following is true for decision tree

A : Decision tree is useful in decision making

B : Decision tree is similar to OLTP

C : Decision Tree is similar to cluster analysis

D : Decision tree needs to find probabilities of hypothesis

Q.no 16. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 17. sensitivity is also known as

A : false rate

B : recall

C : negative rate

D : recognition rate

Q.no 18. Which of the following are methods for supervised classification?

A : Decision tree

B : K-Means

C : Hierarchical

D : Apriori

Q.no 19. The schema is collection of stars. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 20. Removing duplicate records is a process called

A : recovery

B : data cleaning

C : data cleansing

D : data pruning

Q.no 21. The Galaxy Schema is also called as

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 22. Every key structure in the data warehouse contains a time element

A : records

B : Explicitly

C : Implicitly and explicitly

D : Implicitly or explicitly

Q.no 23. If x and y are two objects of nominal attribute with COMP and IT values respectively, then what is the similarity between these two objects?

A : Zero

B : Infinity

C : Two

D : One

Q.no 24. The accuracy of a classifier on a given test set is the percentage of

A : test set tuples that are correctly classified by the classifier

B : test set tuples that are incorrectly classified by the classifier

C : test set tuples that are incorrectly misclassified by the classifier

D : test set tuples that are not classified by the classifier

Q.no 25. A lattice of cuboids is called as

A : Data cube

B : Dimesnion lattice

C : Master lattice

D : Fact table

Q.no 26. What is uniform support in multilevel association rule minig?

A : Use of minimum support

B : Use of minimum support and confidence

C : Use of same minimum threshold at each abstraction level

D : Use of minimum support and support count

Q.no 27. Which of the following is not correct use of cross validation?

A : Selecting variables to include in a model

B : Comparing predictors

C : Selecting parameters in prediction function

D : classification

Q.no 28. The frequent-item-header-table consists of number fields

A : Only one

B : Two

C : Three

D : Four

Q.no 29. Which of these distributions is used for a testing hypothesis?

A : Normal Distribution

B : Chi-Squared Distribution

C : Gamma Distribution

D : Poisson Distribution

Q.no 30. What is the approach of basic algorithm for decision tree induction?

A : Greedy

B : Top Down

C : Procedural

D : Step by Step

Q.no 31. Joins will be needed to execute the query. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 32. Which of the following sequence is used to calculate proximity measures for ordinal attribute?

A : Replacement discretization and distance measure

B : Replacement characterizarion and distance measure

C : Normalization discretization and distance measure

D : Replacement normalization and distance measure

Q.no 33. Some company wants to divide their customers into distinct groups to send offers this is an example of

A : Data Extraction

B : Data Classification

C : Data Discrimination

D : Data Selection

Q.no 34. Which statement is true about the KNN algorithm?

A : All attribute values must be categorical

B : The output attribute must be cateogrical.

C : Attribute values may be either categorical or numeric.

D : All attributes must be numeric.

Q.no 35. The correlation coefficient is used to determine:

A : A specific value of the y-variable given a specific value of the x-variable

B : A specific value of the x-variable given a specific value of the y-variable

C : The strength of the relationship between the x and y variables

D : None of these

Q.no 36. What type of data do you need for a chi-square test?

A : Categorical

B : Ordinal

C : Interval

D : Scales

Q.no 37. In which step of Knowledge Discovery, multiple data sources are combined?

A : Data Cleaning

B : Data Integration

C : Data Selection

D : Data Transformation

Q.no 38. Which of the following is measure of document similarity?

A : Cosine dissimilarity

B : Sine similarity

C : Sine dissimilarity

D : Cosine similarity

Q.no 39. How will you counter over-fitting in decision tree?

A : By creating new rules

B : By pruning the longer rules

C : Both By pruning the longer rules' and ' By creating new rules'

D : BY creating new tree

Q.no 40. In multilevel association rules, which strategy is employed

A : Top-down

B : Recursive

C : Bottom-up

D : Divide and conquer

Q.no 41. precision of model is 0.75 and recall is 0.43 then F-Score is

A : F-Score= 0.99

B : F-Score= 0.84

C : F-Score= 0.55

D : F-Score= 0.49

Q.no 42. Accuracy is

A : Number of correct predictions out of total no. of predictions

B : Number of incorrect predictions out of total no. of predictions

C : Number of predictions out of total no. of predictions

D : Total number of predictions

Q.no 43. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 44. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 45. These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

A : 25

B : 210

C : 62

D : 30

Q.no 46. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 47. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 48. Which one of these is a tree based learner?

A : Rule based

B : Bayesian Belief Network

C : Bayesian classifier

D : Random Forest

Q.no 49. a model predicts 50 examples belonging to the minority class, 45 of which are true positives and five of which are false positives. Precision of model is

A : Precision= 0.90

B : Precision= 0.79

C : Precision= 0.45

D : Precision= 0.68

Q.no 50. Name the property of objects for which distance from first object to second and vice-versa is same.

A : Symmetry

B : Transitive

C : Positive definiteness

D : Traingle inequality

Q.no 51. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 52. Holdout method, Cross-validation and Bootstrap methods are techniques to estimate

A : Precision

B : Classifier performance

C : Recall

D : F-measure

Q.no 53. Ordinal attribute has three distinct values such as Fair, Good, and Excellent.
If x and y are two objects of ordinal attribute with Fair and Good values respectively, then what is the distance from object y to x?

A : 1

B : 0

C : 0.5

D : 0.75

Q.no 54. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 55. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 56. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 57. In Binning, we first sort data and partition into (equal-frequency) bins and then which of the following is not valid step

A : smooth by bin boundaries

B : smooth by bin median

C : smooth by bin means

D : smooth by bin values

Q.no 58. What is the another name of Supremum distance?

A : Wighted Euclidean distance

B : City Block
distance

C : Chebyshev distance

D : Euclidean distance

Q.no 59. In one of the frequent itemset example, it is observed that if tea and milk are bought then sugar is also purchased by customers. After, generating an association rule among the given set of items, it is inferred:

A : {Tea} is antecedent and {sugar} is consequent

B : {Tea} is antecedent and the itemset {milk, sugar} is consequent

C : The itemset {Tea, milk} is consequent and {sugar} is antecedent

D : The itemset { Tea, milk} is antecedent and {sugar} is consequent

Q.no 60. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Answer for Question No 1. is c

Answer for Question No 2. is a

Answer for Question No 3. is a

Answer for Question No 4. is b

Answer for Question No 5. is b

Answer for Question No 6. is b

Answer for Question No 7. is b

Answer for Question No 8. is a

Answer for Question No 9. is d

Answer for Question No 10. is a

Answer for Question No 11. is c

Answer for Question No 12. is c

Answer for Question No 13. is d

Answer for Question No 14. is b

Answer for Question No 15. is a

Answer for Question No 16. is a

Answer for Question No 17. is b

Answer for Question No 18. is a

Answer for Question No 19. is c

Answer for Question No 20. is b

Answer for Question No 21. is c

Answer for Question No 22. is d

Answer for Question No 23. is a

Answer for Question No 24. is a

Answer for Question No 25. is a

Answer for Question No 26. is c

Answer for Question No 27. is d

Answer for Question No 28. is b

Answer for Question No 29. is b

Answer for Question No 30. is a

Answer for Question No 31. is b

Answer for Question No 32. is d

Answer for Question No 33. is b

Answer for Question No 34. is d

Answer for Question No 35. is c

Answer for Question No 36. is a

Answer for Question No 37. is b

Answer for Question No 38. is d

Answer for Question No 39. is b

Answer for Question No 40. is a

Answer for Question No 41. is c

Answer for Question No 42. is a

Answer for Question No 43. is d

Answer for Question No 44. is d

Answer for Question No 45. is d

Answer for Question No 46. is b

Answer for Question No 47. is d

Answer for Question No 48. is d

Answer for Question No 49. is a

Answer for Question No 50. is a

Answer for Question No 51. is b

Answer for Question No 52. is b

Answer for Question No 53. is c

Answer for Question No 54. is a

Answer for Question No 55. is c

Answer for Question No 56. is a

Answer for Question No 57. is d

Answer for Question No 58. is c

Answer for Question No 59. is d

Answer for Question No 60. is d

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. The Synonym for data mining is

A : Data warehouse

B : Knowledge discovery in database

C : ETL

D : Business Intelligemce

Q.no 2. The example of knowledge type constraints in constraint based mining is

A : Association or Correlation

B : Rule templates

C : Task relevant data

D : Threshold measures

Q.no 3. If two documents are similar, then what is the measure of angle between two documents?

A : 30

B : 60

C : 90

D : 0

Q.no 4. The most widely used metrics and tools to assess a classification model are:

A : Conusion Matrix

B : Support

C : Entropy

D : Probability

Q.no 5. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 6. Height is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : Numeric

Q.no 7. How can one represent document to calculate cosine similarity?

A : Vector

B : Matirx

C : List

D : Term frequency vector

Q.no 8. Cotraining is one form of

A : sampling

B : Reinforcement learning

C : unsupervised classification

D : semi-supervised classification

Q.no 9. Which is the keyword that distinguishes data warehouses from other data repository systems ?

A : Subject-oriented

B : Object-oriented

C : Client server

D : Time-invariant

Q.no 10. Self-training is the simplest form of

A : supervised classification

B : semi-supervised classification

C : unsupervised classification

D : regression

Q.no 11. Which of the following is correct about Proximity measures?

A : Similarity

B : Dissimilarity

C : Similarity as well as Dissimilarity

D : Neither similarity nor dissimilarity

Q.no 12. For Apriori algorithm, what is the first phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 13. Hidden knowledge referred to

A : A set of databases from different vendors, possibly using different database paradigms

B : An approach to a problem that is not guaranteed to work but performs well in most cases

C : Information that is hidden in a database and that cannot be recovered by a simple SQL query

D : None of these

Q.no 14. Color is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : numeric

Q.no 15. What is C4.5 is used to build

A : Decision tree

B : Regression Analysis

C : Induction

D : Association Rules

Q.no 16. Choose the correct concept hierarchy.

A : city < street < state < country

B : street < city < state < country

C : street > city > state > country

D : street > city > country > state

Q.no 17. Learning algorithm which trains with combination of labeled and unlabeled data.

A : Supervised

B : Unsupervised

C : Semi supervised

D : Non- supervised

Q.no 18. An automatic car driver and business intelligent systems are examples of

A : Regression

B : Classification

C : Machine Learning

D : Reinforcement learning

Q.no 19. Which of the following is direct application of frequent itemset mining?

A : Social Network Analysis

B : Market Basket Analysis

C : Outlier Detection

D : Intrusion Detection

Q.no 20. recall is a measure of

A : completeness of what percentage
of positive tuples are labeled

B : a measure of exactness for misclassification

C : a measure of exactness of what percentage of tuples are not classified

D : a measure of exactness of what percentage of tuples labeled as
negative are at actual

Q.no 21. The Microsoft SQL Server 2000 is the example of

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 22. Multilevel association rule mining is

A : Association rules generated from candidate-generation method

B : Association rules generated from without candidate-generation method

C : Association rules generated from mining data at multiple abstarction level

D : Assocation rules generated from frequent itemsets

Q.no 23. For mining frequent itemsets, the Data format used by Apriori and FP-Growth algorithms are

A : Apriori uses horizontal and FP-Growth uses vertical data format

B : Apriori uses vertical and FP-Growth uses horizontal data format

C : Apriori and FP-Growth both uses vertical data format

D : Apriori and FP-Growth both uses horizontal data format

Q.no 24. What is uniform support in multilevel association rule minig?

A : Use of minimum support

B : Use of minimum support and confidence

C : Use of same minimum threshold at each abstraction level

D : Use of minimum support and support count

Q.no 25. It is the main technique employed for data selection.

A : Noise

B : Sampling

C : Clustering

D : Histogram

Q.no 26. Where does the bayes rule used?

A : Solving queries

B : Increasing complexity

C : Decreasing complexity

D : Answering probabilistic query

Q.no 27. This operation may add new dimension to the cube

A : Roll up

B : Drill down

C : Slice

D : Dice

Q.no 28. If x and y are two objects of nominal attribute with COMP and IT values respectively, then what is the similarity between these two objects?

A : Zero

B : Infinity

C : Two

D : One

Q.no 29. Every key structure in the data warehouse contains a time element

A : records

B : Explicitly

C : Implicitly and explicitly

D : Implicitly or explicitly

Q.no 30. What type of matrix is required to represent binary data for proximity measures?

A : Normal matrix

B : Sparse matrix

C : Dense matrix

D : Contingency matrix

Q.no 31. In which step of Knowledge Discovery, multiple data sources are combined?

A : Data Cleaning

B : Data Integration

C : Data Selection

D : Data Transformation

Q.no 32. In a decision tree each leaf node represents

A : Test conditions

B : Class labels

C : Attribute values

D : Decision

Q.no 33. Which of the following activities is a data mining task?

A : Monitoring the heart rate of a patient for abnormalities

B : Extracting the frequencies of a sound wave

C : Predicting the outcomes of tossing a (fair) pair of dice

D : Dividing the customers of a company according to their profitability

Q.no 34. To improve the accuracy of multiclass classification we can use

A : cross validation

B : sampling

C : Error-detecting codes

D : Error-correcting codes

Q.no 35. Cross validation involves

A : testing the machine on all possible ways by substituting the original sample into training set

B : testing the machine on all possible ways by dividing the original sample into training and validation sets.

C : testing the machine with only validation sets

D : testing the machine on only testing datasets.

Q.no 36. OLAP Summarization means

A : Consolidated

B : Primitive

C : Highly detailed

D : Recent data

Q.no 37. Identify the example of sequence data

A : weather forecast

B : data matrix

C : market basket data

D : genomic data

Q.no 38. Which of the following is necessary operation to calculate dissimilarity between ordinal attributes?

A : Replacement of ordinal categories

B : Correlation coefficient

C : Discretization

D : Randomization

Q.no 39. How are metarules useful in mining of association rules?

A : Allow users to specify threshold measures

B : Allow users to specify task relevant data

C : Allow users to specify the syntactic forms of rules

D : Allow users to specify correlation or association

Q.no 40. Which of the following probabilities are used in the Bayes theorem.

A : P(Ci|X)

B : P(Ci)

C : P(X|Ci)

D : P(X)

Q.no 41. Ordinal attribute has three distinct values such as Fair, Good, and Excellent.
If x and y are two objects of ordinal attribute with Fair and Good values respectively, then what is the distance from object y to x?

A : 1

B : 0

C : 0.5

D : 0.75

Q.no 42. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 43. Correlation analysis is used for

A : handling missing values

B : identifying redundant attributes

C : handling different data formats

D : eliminating noise

Q.no 44. These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

A : 25

B : 210

C : 62

D : 30

Q.no 45. This technique uses mean and standard deviation scores to transform real-valued attributes.

A : decimal scaling

B : min-max normalization

C : z-score normalization

D : logarithmic normalization

Q.no 46. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 47. Which is the most well known association rule algorithm and is used in most commercial products.

A : Apriori algorithm

B : Pincer-search algorithm

C : Distributed algorithm

D : Partition algorithm

Q.no 48. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 49. Effectiveness of the browsing is highest. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 50. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 51. Name the property of objects for which distance from first object to second and vice-versa is same.

A : Symmetry

B : Transitive

C : Positive definiteness

D : Traingle inequality

Q.no 52. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Q.no 53. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Q.no 54. What is the range of the angle between two term frequency vectors?

A : Zero to Thirty

B : Zero to Ninety

C : Zero to One Eighty

D : Zero to Fourty Five

Q.no 55. precision of model is 0.75 and recall is 0.43 then F-Score is

A : F-Score= 0.99

B : F-Score= 0.84

C : F-Score= 0.55

D : F-Score= 0.49

Q.no 56. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 57. What is the another name of Supremum distance?

A : Wighted Euclidean distance

B : City Block
distance

C : Chebyshev distance

D : Euclidean distance

Q.no 58. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 59. A concept hierarchy that is a total or partial order among attributes in a database schema is called

A : Mixed hierarchy

B : Total hierarchy

C : Schema hierarchy

D : Concept generalization

Q.no 60. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Answer for Question No 1. is b

Answer for Question No 2. is a

Answer for Question No 3. is d

Answer for Question No 4. is a

Answer for Question No 5. is b

Answer for Question No 6. is d

Answer for Question No 7. is d

Answer for Question No 8. is d

Answer for Question No 9. is a

Answer for Question No 10. is b

Answer for Question No 11. is c

Answer for Question No 12. is c

Answer for Question No 13. is c

Answer for Question No 14. is a

Answer for Question No 15. is a

Answer for Question No 16. is b

Answer for Question No 17. is c

Answer for Question No 18. is d

Answer for Question No 19. is b

Answer for Question No 20. is a

Answer for Question No 21. is c

Answer for Question No 22. is c

Answer for Question No 23. is d

Answer for Question No 24. is c

Answer for Question No 25. is b

Answer for Question No 26. is d

Answer for Question No 27. is b

Answer for Question No 28. is a

Answer for Question No 29. is d

Answer for Question No 30. is d

Answer for Question No 31. is b

Answer for Question No 32. is b

Answer for Question No 33. is a

Answer for Question No 34. is d

Answer for Question No 35. is c

Answer for Question No 36. is a

Answer for Question No 37. is d

Answer for Question No 38. is a

Answer for Question No 39. is c

Answer for Question No 40. is a

Answer for Question No 41. is c

Answer for Question No 42. is d

Answer for Question No 43. is b

Answer for Question No 44. is d

Answer for Question No 45. is c

Answer for Question No 46. is a

Answer for Question No 47. is a

Answer for Question No 48. is a

Answer for Question No 49. is a

Answer for Question No 50. is d

Answer for Question No 51. is a

Answer for Question No 52. is d

Answer for Question No 53. is b

Answer for Question No 54. is b

Answer for Question No 55. is c

Answer for Question No 56. is b

Answer for Question No 57. is c

Answer for Question No 58. is a

Answer for Question No 59. is c

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. Which one of the following is true for decision tree

A : Decision tree is useful in decision making

B : Decision tree is similar to OLTP

C : Decision Tree is similar to cluster analysis

D : Decision tree needs to find probabilities of hypothesis

Q.no 2. The first steps involved in the knowledge discovery is?

A : Data Integration

B : Data Selection

C : Data Transformation

D : Data Cleaning

Q.no 3. What is C4.5 is used to build

A : Decision tree

B : Regression Analysis

C : Induction

D : Association Rules

Q.no 4. Which of the following is not frequent pattern?

A : Itemsets

B : Subsequences

C : Substructures

D : Associations

Q.no 5. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 6. A data cube is defined by

A : Dimensions

B : Facts

C : Dimensions and Facts

D : Dimensions or Facts

Q.no 7. An ROC curve for a given model shows the trade-off between

A : random sampling

B : test data and train data

C : cross validation

D : the true positive rate (TPR) and the false positive rate (FPR)

Q.no 8. Which of the following is the data mining tool?

A : Borland C

B : Weka

C : Borland C++

D : Visual C

Q.no 9. Cotraining is one form of

A : sampling

B : Reinforcement learning

C : unsupervised classification

D : semi-supervised classification

Q.no 10. Each dimension is represented by only one table. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 11. What are two steps of tree pruning work?

A : Pessimistic pruning and Optimistic pruning

B : Postpruning and Prepruning

C : Cost complexity pruning and time complexity pruning

D : None of the options

Q.no 12. What do you mean by dissimilarity measure of two objects?

A : Is a numerical measure of how alike two data objects are.

B : Is a numerical measure of how different two data objects are.

C : Higher when objects are more alike

D : Lower when objects are more different

Q.no 13. Choose the correct concept hierarchy.

A : city < street < state < country

B : street < city < state < country

C : street > city > state > country

D : street > city > country > state

Q.no 14. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 15. to evaluate a classifier's quality we use

A : confusion matrix

B : error detection code

C : error correction code

D : classifier

Q.no 16. accuracy is used to measure

A : classifier's true abilities

B : classifier's analytic abilities

C : classifier's decision abilities

D : classifier's predictive abilities

Q.no 17. Supervised learning and unsupervised clustering both require at least one

A : hidden attribute

B : output attribute

C : input attribute

D : categorical attribute

Q.no 18. CART stands for

A : Regression

B : Classification

C : Classification and Regression Trees

D : Decision Trees

Q.no 19. What are closed frequent itemsets?

A : A closed itemset

B : A frequent itemset

C : An itemset which is both closed and frequent

D : Not frequent itemset

Q.no 20. In Data Characterization, class under study is called as?

A : Study Class

B : Intial Class

C : Target Class

D : Final Class

Q.no 21. A nearest neighbor approach is best used

A : with large-sized datasets.

B : when irrelevant attributes have been removed from the data.

C : when a generalized model of the data is desireable.

D : when an explanation of what has been found is of primary importance.

Q.no 22. Lazy learner classification approach is

A : learner waits until the last minute before constructing model to classify

B : a given training data constructs a model first and then uses it to classify

C : the network is constructed by human experts

D : None of the options

Q.no 23. Which of the following probabilities are used in the Bayes theorem.

A : P(Ci|X)

B : P(Ci)

C : P(X|Ci)

D : P(X)

Q.no 24. A frequent pattern tree is a tree structure consisting of

A : A frequent-item-node

B : An item-prefix-tree

C : A frequent-item-header table

D : both B and C

Q.no 25. Holdout and random subsampling are common techniques for assessing

A : K-Fold validation

B : cross validation

C : accuracy

D : sampling

Q.no 26. Specificity is also referred to as

A : true negative rate

B : correctness

C : misclassification rate

D : True positive rate

Q.no 27. If A, B are two sets of items, and A is a subset of B. Which of the following statement is always true?

A : Support(A) is less than or equal to Support(B)

B : Support(A) is greater than or equal to Support(B)

C : Support(A) is equal to Support(B)

D : Support(A) is not equal to Support(B)

Q.no 28. To improve the accuracy of multiclass classification we can use

A : cross validation

B : sampling

C : Error-detecting codes

D : Error-correcting codes

Q.no 29. What is the limitation behind rule generation in Apriori algorithm?

A : Need to generate a huge number of candidate sets

B : Need to repeatedly scan the whole database and Check a large set of candidates by pattern matching

C : Dropping itemsets with valued information

D : Both (a) dnd (b)

Q.no 30. one-versus-one(OVO) and one-versus-all (OVA) classification involves

A : more than two classes

B : Only two classes

C : Only one class

D : No class

Q.no 31. OLAP Summarization means

A : Consolidated

B : Primitive

C : Highly detailed

D : Recent data

Q.no 32. When you use cross validation in machine learning, it means

A : you verify how accurate your model is on multiple and different subsets of data.

B : you verify how accurate your model is on same dataset.

C : you verify how accurate your model is on new dataset.

D : you verify how accurate your model on unknown dataset

Q.no 33. What is the approach of basic algorithm for decision tree induction?

A : Greedy

B : Top Down

C : Procedural

D : Step by Step

Q.no 34. Which of the following operations are used to calculate proximity measures for ordinal attribute?

A : Replacement and discretization

B : Replacement and characterizarion

C : Replacement and normalization

D : Normalization and discretization

Q.no 35. In Apriori algorithm, for generating e. g. 5 itemsets, we use

A : Frequent 5 itemsets

B : Frequent 3 itemsets

C : Frequent 4 itemsets

D : Frequent 6 itemsets

Q.no 36. Which of the following is a predictive model?

A : Clustering

B : Regression

C : Summarization

D : Association rules

Q.no 37. It is the main technique employed for data selection.

A : Noise

B : Sampling

C : Clustering

D : Histogram

Q.no 38. Some company wants to divide their customers into distinct groups to send offers this is an example of

A : Data Extraction

B : Data Classification

C : Data Discrimination

D : Data Selection

Q.no 39. In asymmetric attribute

A : No value is considered important over other values

B : All values are equal

C : Only non-zero value is important

D : Range of values is important

Q.no 40. A lattice of cuboids is called as

A : Data cube

B : Dimesnion lattice

C : Master lattice

D : Fact table

Q.no 41. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 42. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 43. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 44. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 45. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 46. Which operation data warehouse requires ?

A : Initial loading of data

B : Transaction processing

C : Recovery

D : Concurrency control mechanisms

Q.no 47. These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

A : 25

B : 210

C : 62

D : 30

Q.no 48. If True Positives (TP): 7, False Positives (FP): 1,False Negatives (FN): 4, True Negatives (TN): 18. Calculate Precision and Recall.

A : Precision = 0.88, Recall=0.64

B : Precision = 0.44, Recall=0.78

C : Precision = 0.88, Recall=0.22

D : Precision = 0.77, Recall=0.55

Q.no 49. The problem of finding hidden structure from unlabeled data is called as

A : Supervised learning

B : Unsupervised learning

C : Reinforcement Learning

D : Semisupervised learning

Q.no 50. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 51. High entropy means that the partitions in classification are

A : pure

B : Not pure

C : Useful

D : Not useful

Q.no 52. A model makes predictions and predicts 120 examples as belonging to the minority class, 90 of which are correct, and 30 of which are incorrect. Precision of model is

A : Precision = 0.89

B : Precision = 0.23

C : Precision = 0.45

D : Precision = 0.75

Q.no 53. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 54. precision of model is 0.75 and recall is 0.43 then F-Score is

A : F-Score= 0.99

B : F-Score= 0.84

C : F-Score= 0.55

D : F-Score= 0.49

Q.no 55. A model makes predictions and predicts 90 of the positive class predictions correctly and 10 incorrectly.Recall of model is

A : Recall=0.9

B : Recall=0.39

C : Recall=0.65

D : Recall=5.0

Q.no 56. Effectiveness of the browsing is highest. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 57. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 58. A data normalization technique for real-valued attributes that divides each numerical value by the same power of 10.

A : min-max normalization

B : z-score normalization

C : decimal scaling

D : decimal smoothing

Q.no 59. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 60. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Answer for Question No 1. is a

Answer for Question No 2. is d

Answer for Question No 3. is a

Answer for Question No 4. is d

Answer for Question No 5. is b

Answer for Question No 6. is c

Answer for Question No 7. is d

Answer for Question No 8. is b

Answer for Question No 9. is d

Answer for Question No 10. is a

Answer for Question No 11. is b

Answer for Question No 12. is b

Answer for Question No 13. is b

Answer for Question No 14. is a

Answer for Question No 15. is a

Answer for Question No 16. is d

Answer for Question No 17. is c

Answer for Question No 18. is c

Answer for Question No 19. is c

Answer for Question No 20. is c

Answer for Question No 21. is b

Answer for Question No 22. is a

Answer for Question No 23. is a

Answer for Question No 24. is d

Answer for Question No 25. is c

Answer for Question No 26. is a

Answer for Question No 27. is b

Answer for Question No 28. is d

Answer for Question No 29. is d

Answer for Question No 30. is a

Answer for Question No 31. is a

Answer for Question No 32. is a

Answer for Question No 33. is a

Answer for Question No 34. is c

Answer for Question No 35. is c

Answer for Question No 36. is b

Answer for Question No 37. is b

Answer for Question No 38. is b

Answer for Question No 39. is c

Answer for Question No 40. is a

Answer for Question No 41. is a

Answer for Question No 42. is d

Answer for Question No 43. is c

Answer for Question No 44. is b

Answer for Question No 45. is a

Answer for Question No 46. is a

Answer for Question No 47. is d

Answer for Question No 48. is a

Answer for Question No 49. is b

Answer for Question No 50. is a

Answer for Question No 51. is b

Answer for Question No 52. is d

Answer for Question No 53. is b

Answer for Question No 54. is c

Answer for Question No 55. is a

Answer for Question No 56. is a

Answer for Question No 57. is a

Answer for Question No 58. is c

Answer for Question No 59. is d

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. For Apriori algorithm, what is the second phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 2. Which of these is not a frequent pattern mining algorithm?

A : Decision trees

B : Eclat

C : FP growth

D : Apriori

Q.no 3. Which of the following is not a type of constraints?

A : Data constraints

B : Rule constraints

C : Knowledge type constraints

D : Time constraints

Q.no 4. An ROC curve for a given model shows the trade-off between

A : random sampling

B : test data and train data

C : cross validation

D : the true positive rate (TPR) and the false positive rate (FPR)

Q.no 5. If two documents are similar, then what is the measure of angle between two documents?

A : 30

B : 60

C : 90

D : 0

Q.no 6. Choose the correct concept hierarchy.

A : city < street < state < country

B : street < city < state < country

C : street > city > state > country

D : street > city > country > state

Q.no 7. Supervised learning and unsupervised clustering both require at least one

A : hidden attribute

B : output attribute

C : input attribute

D : categorical attribute

Q.no 8. The fact is also called as

A : Dimension

B : Key

C : Schema

D : Measure

Q.no 9. The most widely used metrics and tools to assess a classification model are:

A : Conusion Matrix

B : Support

C : Entropy

D : Probability

Q.no 10. A person trained to interact with a human expert in order to capture their knowledge.

A : knowledge programmer

B : knowledge developer

C : knowledge engineer

D : knowledge extractor

Q.no 11. Training process that generates tree is called as

A : Pruning

B : Rule generation

C : Induction

D : spliiting

Q.no 12. The schema is collection of stars. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 13. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 14. to evaluate a classifier's quality we use

A : confusion matrix

B : error detection code

C : error correction code

D : classifier

Q.no 15. For Apriori algorithm, what is the first phase?

A : Pruning

B : Partitioning

C : Candidate generation

D : Itemset generation

Q.no 16. The example of knowledge type constraints in constraint based mining is

A : Association or Correlation

B : Rule templates

C : Task relevant data

D : Threshold measures

Q.no 17. Height is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : Numeric

Q.no 18. A data cube is defined by

A : Dimensions

B : Facts

C : Dimensions and Facts

D : Dimensions or Facts

Q.no 19. Which one of the following is true for decision tree

A : Decision tree is useful in decision making

B : Decision tree is similar to OLTP

C : Decision Tree is similar to cluster analysis

D : Decision tree needs to find probabilities of hypothesis

Q.no 20. What are two steps of tree pruning work?

A : Pessimistic pruning and Optimistic pruning

B : Postpruning and Prepruning

C : Cost complexity pruning and time complexity pruning

D : None of the options

Q.no 21. The Microsoft SQL Server 2000 is the example of

A : ROLAP

B : MOLAP

C : HOLAP

D : HaoLap

Q.no 22. The property of Apriori algorithm is

A : All nonempty subsets of a frequent itemsets must also be frequent

B : All empty subsets of a frequent itemsets must also be frequent

C : All nonempty subsets of a frequent itemsets must be not frequent

D : All nonempty subsets of a frequent itemsets can frequent or not frequent

Q.no 23. Multilevel association rule mining is

A : Association rules generated from candidate-generation method

B : Association rules generated from without candidate-generation method

C : Association rules generated from mining data at multiple abstarction level

D : Assocation rules generated from frequent itemsets

Q.no 24. Which of the following activities is a data mining task?

A : Monitoring the heart rate of a patient for abnormalities

B : Extracting the frequencies of a sound wave

C : Predicting the outcomes of tossing a (fair) pair of dice

D : Dividing the customers of a company according to their profitability

Q.no 25. What type of matrix is required to represent binary data for proximity measures?

A : Normal matrix

B : Sparse matrix

C : Dense matrix

D : Contingency matrix

Q.no 26. Sensitivity is also referred to as

A : misclassification rate

B : true negative rate

C : True positive rate

D : correctness

Q.no 27. In Apriori algorithm, for generating e. g. 5 itemsets, we use

A : Frequent 5 itemsets

B : Frequent 3 itemsets

C : Frequent 4 itemsets

D : Frequent 6 itemsets

Q.no 28. Handwritten digit recognition classifying an image of a handwritten number into a digit from 0 to 9 is example of

A : Multiclassification

B : Multi-label classification

C : Imbalanced classification

D : Binary Classification

Q.no 29. A lattice of cuboids is called as

A : Data cube

B : Dimesnion lattice

C : Master lattice

D : Fact table

Q.no 30. Specificity is also referred to as

A : true negative rate

B : correctness

C : misclassification rate

D : True positive rate

Q.no 31. To improve the accuracy of multiclass classification we can use

A : cross validation

B : sampling

C : Error-detecting codes

D : Error-correcting codes

Q.no 32. This operation may add new dimension to the cube

A : Roll up

B : Drill down

C : Slice

D : Dice

Q.no 33. The Galaxy Schema is also called as

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 34. For a classification problem with highly imbalanced class. The majority class is observed 99% of times in the training data.
Your model has 99% accuracy after taking the predictions on test data. Which of the following is not true in such a case?

A : Imbalaced problems should not be measured using Accuracy metric.

B : Accuracy metric is not a good idea for imbalanced class problems.

C : Precision and recall metrics aren't good for imbalanced class problems.

D : Precision and recall metrics are good for imbalanced class problems.

Q.no 35. one-versus-one(OVO) and one-versus-all (OVA) classification involves

A : more than two classes

B : Only two classes

C : Only one class

D : No class

Q.no 36. How are metarules useful in mining of association rules?

A : Allow users to specify threshold measures

B : Allow users to specify task relevant data

C : Allow users to specify the syntactic forms of rules

D : Allow users to specify correlation or association

Q.no 37. OLAP Summarization means

A : Consolidated

B : Primitive

C : Highly detailed

D : Recent data

Q.no 38. A frequent pattern tree is a tree structure consisting of

A : A frequent-item-node

B : An item-prefix-tree

C : A frequent-item-header table

D : both B and C

Q.no 39. The confusion matrix is a useful tool for analyzing

A : Regression

B : Classification

C : Sampling

D : Cross validation

Q.no 40. Cross validation involves

A : testing the machine on all possible ways by substituting the original sample into training set

B : testing the machine on all possible ways by dividing the original sample into training and validation sets.

C : testing the machine with only validation sets

D : testing the machine on only testing datasets.

Q.no 41. Which one of these is a tree based learner?

A : Rule based

B : Bayesian Belief Network

C : Bayesian classifier

D : Random Forest

Q.no 42. Ordinal attribute has three distinct values such as Fair, Good, and Excellent.
If x and y are two objects of ordinal attribute with Fair and Good values respectively, then what is the distance from object y to x?

A : 1

B : 0

C : 0.5

D : 0.75

Q.no 43. Rotating the axes in a 3-D cube is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 44. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 45. If True Positives (TP): 7, False Positives (FP): 1,False Negatives (FN): 4, True Negatives (TN): 18. Calculate Precision and Recall.

A : Precision = 0.88, Recall=0.64

B : Precision = 0.44, Recall=0.78

C : Precision = 0.88, Recall=0.22

D : Precision = 0.77, Recall=0.55

Q.no 46. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 47. Accuracy is

A : Number of correct predictions out of total no. of predictions

B : Number of incorrect predictions out of total no. of predictions

C : Number of predictions out of total no. of predictions

D : Total number of predictions

Q.no 48. What is the range of the angle between two term frequency vectors?

A : Zero to Thirty

B : Zero to Ninety

C : Zero to One Eighty

D : Zero to Fourty Five

Q.no 49. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 50. Transforming a 3-D cube into a series of 2-D planes is the examplele of

A : Pivot

B : Roll up

C : Drill down

D : Slice

Q.no 51. A model makes predictions and predicts 120 examples as belonging to the minority class, 90 of which are correct, and 30 of which are incorrect. Precision of model is

A : Precision = 0.89

B : Precision = 0.23

C : Precision = 0.45

D : Precision = 0.75

Q.no 52. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 53. A data normalization technique for real-valued attributes that divides each numerical value by the same power of 10.

A : min-max normalization

B : z-score normalization

C : decimal scaling

D : decimal smoothing

Q.no 54. High entropy means that the partitions in classification are

A : pure

B : Not pure

C : Useful

D : Not useful

Q.no 55. In Binning, we first sort data and partition into (equal-frequency) bins and then which of the following is not valid step

A : smooth by bin boundaries

B : smooth by bin median

C : smooth by bin means

D : smooth by bin values

Q.no 56. This technique uses mean and standard deviation scores to transform real-valued attributes.

A : decimal scaling

B : min-max normalization

C : z-score normalization

D : logarithmic normalization

Q.no 57. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 58. precision of model is 0.75 and recall is 0.43 then F-Score is

A : F-Score= 0.99

B : F-Score= 0.84

C : F-Score= 0.55

D : F-Score= 0.49

Q.no 59. The basic idea of the apriori algorithm is to generate the item sets of a particular size & scans the database. These item sets are

A : Primary

B : Secondary

C : Superkey

D : Candidate

Q.no 60. How the bayesian network can be used to answer any query?

A : Full distribution

B : Joint distribution

C : Partial distribution

D : All of the mentioned

Answer for Question No 1. is a

Answer for Question No 2. is a

Answer for Question No 3. is d

Answer for Question No 4. is d

Answer for Question No 5. is d

Answer for Question No 6. is b

Answer for Question No 7. is c

Answer for Question No 8. is d

Answer for Question No 9. is a

Answer for Question No 10. is c

Answer for Question No 11. is c

Answer for Question No 12. is c

Answer for Question No 13. is b

Answer for Question No 14. is a

Answer for Question No 15. is c

Answer for Question No 16. is a

Answer for Question No 17. is d

Answer for Question No 18. is c

Answer for Question No 19. is a

Answer for Question No 20. is b

Answer for Question No 21. is c

Answer for Question No 22. is a

Answer for Question No 23. is c

Answer for Question No 24. is a

Answer for Question No 25. is d

Answer for Question No 26. is c

Answer for Question No 27. is c

Answer for Question No 28. is a

Answer for Question No 29. is a

Answer for Question No 30. is a

Answer for Question No 31. is d

Answer for Question No 32. is b

Answer for Question No 33. is c

Answer for Question No 34. is c

Answer for Question No 35. is a

Answer for Question No 36. is c

Answer for Question No 37. is a

Answer for Question No 38. is d

Answer for Question No 39. is b

Answer for Question No 40. is c

Answer for Question No 41. is d

Answer for Question No 42. is c

Answer for Question No 43. is a

Answer for Question No 44. is b

Answer for Question No 45. is a

Answer for Question No 46. is b

Answer for Question No 47. is a

Answer for Question No 48. is b

Answer for Question No 49. is d

Answer for Question No 50. is a

Answer for Question No 51. is d

Answer for Question No 52. is c

Answer for Question No 53. is c

Answer for Question No 54. is b

Answer for Question No 55. is d

Answer for Question No 56. is c

Answer for Question No 57. is d

Answer for Question No 58. is c

Answer for Question No 59. is d

Answer for Question No 60. is b

Seat No -

Total number of questions : 60

12695_Data Mining and Warehousing

Time : 1hr

Max Marks : 50

N.B

1) All questions are Multiple Choice Questions having single correct option.

2) Attempt any 50 questions out of 60.

3) Use of calculator is allowed.

4) Each question carries 1 Mark.

5) Specially abled students are allowed 20 minutes extra for examination.

6) Do not use pencils to darken answer.

7) Use only black/blue ball point pen to darken the appropriate circle.

8) No change will be allowed once the answer is marked on OMR Sheet.

9) Rough work shall not be done on OMR sheet or on question paper.

10) Darken ONLY ONE CIRCLE for each answer.

Q.no 1. What is the method to interpret the results after rule generation?

A : Absolute Mean

B : Lift ratio

C : Gini Index

D : Apriori

Q.no 2. OLAP database design is

A : Application-oriented

B : Object-oriented

C : Goal-oriented

D : Subject-oriented

Q.no 3. Multilevel association rules can be mined efficiently using

A : Support

B : Confidence

C : Support count

D : Concept Hierarchies under support-confidence framework

Q.no 4. accuracy is used to measure

A : classifier’s true abilities

B : classifier’s analytic abilities

C : classifier’s decision abilities

D : classifier’s predictive abilities

Q.no 5. Supervised learning and unsupervised clustering both require at least one

A : hidden attribute

B : output attribute

C : input attribute

D : categorical attribute

Q.no 6. The task of building decision model from labeled training data is called as

A : Supervised Learning

B : Unsupervised Learning

C : Reinforcement Learning

D : Structure Learning

Q.no 7. What is the range of the cosine similarity of the two documents?

A : Zero to One

B : Zero to infinity

C : Infinity to infinity

D : Zero to Zero

Q.no 8. Multi-class classification makes the assumption that each sample is assigned to

A : one and only one label

B : many labels

C : one or many labels

D : no label

Q.no 9. Which of these is not a frequent pattern mining algorithm?

A : Decision trees

B : Eclat

C : FP growth

D : Apriori

Q.no 10. The first steps involved in the knowledge discovery is?

A : Data Integration

B : Data Selection

C : Data Transformation

D : Data Cleaning

Q.no 11. The distance between two points calculated using Pythagoras theorem is

A : Supremum distance

B : Euclidean distance

C : Linear distance

D : Manhattan Distance

Q.no 12. What do you mean by dissimilarity measure of two objects?

A : Is a numerical measure of how alike two data objects are.

B : Is a numerical measure of how different two data objects are.

C : Higher when objects are more alike

D : Lower when objects are more different

Q.no 13. An ROC curve for a given model shows the trade-off between

A : random sampling

B : test data and train data

C : cross validation

D : the true positive rate (TPR) and the false positive rate (FPR)

Q.no 14. Each dimension is represented by only one table. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 15. Choose the correct concept hierarchy.

A : city < street < state < country

B : street < city < state < country

C : street > city > state > country

D : street > city > country > state

Q.no 16. Height is an example of which type of attribute

A : Nominal

B : Binary

C : Ordinal

D : Numeric

Q.no 17. Which angle is used to measure document similarity?

A : Sin

B : Tan

C : Cos

D : Sec

Q.no 18. Which of the following is the data mining tool?

A : Borland C

B : Weka

C : Borland C++

D : Visual C

Q.no 19. A decision tree is also known as

A : general tree

B : binary tree

C : prediction tree

D : None of the options

Q.no 20. recall is a measure of

A : completeness of what percentage
of positive tuples are labeled

B : a measure of exactness for misclassification

C : a measure of exactness of what percentage of tuples are not classified

D : a measure of exactness of what percentage of tuples labeled as
negative are at actual

Q.no 21. What is the approach of basic algorithm for decision tree induction?

A : Greedy

B : Top Down

C : Procedural

D : Step by Step

Q.no 22. The rule is considered as intersting if

A : They satisfy both minimum support and minimum confidence threshold

B : They satisfy both maximum support and maximum confidence threshold

C : They satisfy maximum support and minimum confidence threshold

D : They satisfy minimum support and maximum confidence threshold

Q.no 23. For mining frequent itemsets, the Data format used by Apriori and FP-Growth algorithms are

A : Apriori uses horizontal and FP-Growth uses vertical data format

B : Apriori uses vertical and FP-Growth uses horizontal data format

C : Apriori and FP-Growth both uses vertical data format

D : Apriori and FP-Growth both uses horizontal data format

Q.no 24. Which of the following sequence is used to calculate proximity measures for ordinal attribute?

A : Replacement discretization and distance measure

B : Replacement characterizarion and distance measure

C : Normalization discretization and distance measure

D : Replacement normalization and distance measure

Q.no 25. Multilevel association rule mining is

A : Association rules generated from candidate-generation method

B : Association rules generated from without candidate-generation method

C : Association rules generated from mining data at multiple abstarction level

D : Assocation rules generated from frequent itemsets

Q.no 26. Which of the following is not correct use of cross validation?

A : Selecting variables to include in a model

B : Comparing predictors

C : Selecting parameters in prediction function

D : classification

Q.no 27. What do you mean by support(A)?

A : Total number of transactions containing A

B : Total Number of transactions not containing A

C : Number of transactions containing A / Total number of transactions

D : Number of transactions not containing A / Total number of transactions

Q.no 28. The fact table contains

A : The names of the facts

B : Keys to each of the related dimension tables

C : Facts and keys

D : Facts or keys

Q.no 29. Every key structure in the data warehouse contains a time element

A : records

B : Explicitly

C : Implicitly and explicitly

D : Implicitly or explicitly

Q.no 30. The accuracy of a classifier on a given test set is the percentage of

A : test set tuples that are correctly classified by the classifier

B : test set tuples that are incorrectly classified by the classifier

C : test set tuples that are incorrectly misclassified by the classifier

D : test set tuples that are not classified by the classifier

Q.no 31. How will you counter over-fitting in decision tree?

A : By creating new rules

B : By pruning the longer rules

C : Both By pruning the longer rules' and ' By creating new rules'

D : BY creating new tree

Q.no 32. The confusion matrix is a useful tool for analyzing

A : Regression

B : Classification

C : Sampling

D : Cross validation

Q.no 33. If A, B are two sets of items, and A is a subset of B. Which of the following statement is always true?

A : Support(A) is less than or equal to Support(B)

B : Support(A) is greater than or equal to Support(B)

C : Support(A) is equal to Support(B)

D : Support(A) is not equal to Support(B)

Q.no 34. What is the limitation behind rule generation in Apriori algorithm?

A : Need to generate a huge number of candidate sets

B : Need to repeatedly scan the whole database and Check a large set of candidates by pattern matching

C : Dropping itemsets with valued information

D : Both (a) dnd (b)

Q.no 35. In asymmetric attribute

A : No value is considered important over other values

B : All values are equal

C : Only non-zero value is important

D : Range of values is important

Q.no 36. One of the most well known software used for classification is

A : Java

B : C4.5

C : Oracle

D : C++

Q.no 37. Identify the example of sequence data

A : weather forecast

B : data matrix

C : market basket data

D : genomic data

Q.no 38. What type of matrix is required to represent binary data for proximity measures?

A : Normal matrix

B : Sparse matrix

C : Dense matrix

D : Contingency matrix

Q.no 39. Some company wants to divide their customers into distinct groups to send offers this is an example of

A : Data Extraction

B : Data Classification

C : Data Discrimination

D : Data Selection

Q.no 40. This operation may add new dimension to the cube

A : Roll up

B : Drill down

C : Slice

D : Dice

Q.no 41. Which of the following sentence is FALSE regarding regression?

A : It relates inputs to outputs.

B : It is used for prediction.

C : It may be used for interpretation.

D : It discovers causal relationships.

Q.no 42. The following represents age distribution of students in an elementary class. Find the mode of the values: 7, 9, 10, 13, 11, 7, 9, 19, 12, 11, 9, 7, 9, 10, 11.

A : 7

B : 9

C : 10

D : 11

Q.no 43. These numbers are taken from the number of people that attended a particular church every Friday for 7 weeks: 62, 18, 39, 13, 16, 37, 25. Find the mean.

A : 25

B : 210

C : 62

D : 30

Q.no 44. Effectiveness of the browsing is highest. Recognize the type of schema.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 45. The cuboid that holds the lowest level of summarization is called as

A : 0-D cuboid

B : 1-D cuboid

C : Base cuboid

D : 2-D cuboid

Q.no 46. The tables are easy to maintain and saves storage space.

A : Star Schema

B : Snowflake schema

C : Fact constellation

D : Database schema

Q.no 47. A model makes predictions and predicts 120 examples as belonging to the minority class, 90 of which are correct, and 30 of which are incorrect. Precision of model is

A : Precision = 0.89

B : Precision = 0.23

C : Precision = 0.45

D : Precision = 0.75

Q.no 48. A database has 4 transactions.Of these, 4 transactions include milk and bread. Further , of the given 4 transactions, 3 transactions include cheese. Find the support percentage for the following association rule, " If milk and bread purchased then cheese is also purchased".

A : 0.6

B : 0.75

C : 0.8

D : 0.7

Q.no 49. What is the range of the angle between two term frequency vectors?

A : Zero to Thirty

B : Zero to Ninety

C : Zero to One Eighty

D : Zero to Fourty Five

Q.no 50. What does a Pearson's product-moment allow you to identify?

A : Whether there is a relationship between variables

B : Whether there is a significant effect and interaction of independent variables

C : Whether there is a significant difference between variables

D : Whether there is a significant effect and interaction of dependent variables

Q.no 51. Consider three itemsets V1={tomato, potato,onion}, V2={tomato,potato}, V3={tomato}. Which of the following statement is correct?

A : support(V1) is greater than support (V2)

B : support(V3) is greater than support (V2)

C : support(V1) is greater than support(V3)

D : support(V2) is greater than support(V3)

Q.no 52. What is the another name of Supremum distance?

A : Wighted Euclidean distance

B : City Block
distance

C : Chebyshev distance

D : Euclidean distance

Q.no 53. This technique uses mean and standard deviation scores to transform real-valued attributes.

A : decimal scaling

B : min-max normalization

C : z-score normalization

D : logarithmic normalization

Q.no 54. When do we use Manhattan distance in data mining?

A : Dimension of the data decreases

B : Dimension of the data increases

C : Underfitting

D : Moderate size of the dimensions

Q.no 55. Correlation analysis is used for

A : handling missing values

B : identifying redundant attributes

C : handling different data formats

D : eliminating noise

Q.no 56. If True Positives (TP): 7, False Positives (FP): 1,False Negatives (FN): 4, True Negatives (TN): 18. Calculate Precision and Recall.

A : Precision = 0.88, Recall=0.64

B : Precision = 0.44, Recall=0.78

C : Precision = 0.88, Recall=0.22

D : Precision = 0.77, Recall=0.55

Q.no 57. A sub-database which consists of set of prefix paths in the FP-tree co-occuring with the sufix pattern is called as

A : Suffix path

B : FP-tree

C : Prefix path

D : Condition pattern base

Q.no 58. Cost complexity pruning algorithm is used in?

A : CART

B : C4.5

C : ID3

D : ALL

Q.no 59. Which is the most well known association rule algorithm and is used in most commercial products.

A : Apriori algorithm

B : Pincer-search algorithm

C : Distributed algorithm

D : Partition algorithm

Q.no 60. Which operation is required to calculate Hamming distacne between two objects?

A : AND

B : OR

C : NOT

D : XOR

Answer for Question No 1. is b

Answer for Question No 2. is d

Answer for Question No 3. is d

Answer for Question No 4. is d

Answer for Question No 5. is c

Answer for Question No 6. is a

Answer for Question No 7. is a

Answer for Question No 8. is a

Answer for Question No 9. is a

Answer for Question No 10. is d

Answer for Question No 11. is b

Answer for Question No 12. is b

Answer for Question No 13. is d

Answer for Question No 14. is a

Answer for Question No 15. is b

Answer for Question No 16. is d

Answer for Question No 17. is c

Answer for Question No 18. is b

Answer for Question No 19. is c

Answer for Question No 20. is a

Answer for Question No 21. is a

Answer for Question No 22. is a

Answer for Question No 23. is d

Answer for Question No 24. is d

Answer for Question No 25. is c

Answer for Question No 26. is d

Answer for Question No 27. is c

Answer for Question No 28. is c

Answer for Question No 29. is d

Answer for Question No 30. is a

Answer for Question No 31. is b

Answer for Question No 32. is b

Answer for Question No 33. is b

Answer for Question No 34. is d

Answer for Question No 35. is c

Answer for Question No 36. is b

Answer for Question No 37. is d

Answer for Question No 38. is d

Answer for Question No 39. is b

Answer for Question No 40. is b

Answer for Question No 41. is d

Answer for Question No 42. is b

Answer for Question No 43. is d

Answer for Question No 44. is a

Answer for Question No 45. is c

Answer for Question No 46. is b

Answer for Question No 47. is d

Answer for Question No 48. is a

Answer for Question No 49. is b

Answer for Question No 50. is a

Answer for Question No 51. is b

Answer for Question No 52. is c

Answer for Question No 53. is c

Answer for Question No 54. is b

Answer for Question No 55. is b

Answer for Question No 56. is a

Answer for Question No 57. is d

Answer for Question No 58. is a

Answer for Question No 59. is a

Answer for Question No 60. is d