



HALDIA INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION UNDER MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY, WEST BENGAL)

Paper Code: PEC-CS 602B

Paper Name: DATA MINING

Time Allotted: 2 Hours

Full Marks: 70

*The figures in the margin indicate full marks
Candidates are required to give their answers in their own words as far as practicable*

Group – A

(Multiple Choice Type Questions)

1. Choose the correct alternatives from the followings:

30 x 1 = 30

(i) Removal of noisy and irrelevant data is known as [CO1]

a) Data Cleaning b) Data Aggregation c) Extraction of Data d) Compression of Data

(ii) Identify the incorrect option among the following which is not involved in data mining. [CO1]

a) Data Exploration b) Knowledge Extraction
c) Data Transformation d) Data archaeology

(iii) Identify the options below that a data warehouse can include [CO2]

a) Database Table b) Online Data c) Flat Files d) All of the above

(iv) Choose the incorrect property of the data warehouse. [CO2]

a) Collection from heterogeneous sources b) Subject oriented
c) Time variant d) Volatile

(v) Which is not the part of KDD process from the following? [CO1]

a) Selection b) Pre-processing c) Reduction d) Summation

(vi) Data Mining turns a large collection of data into [CO1]

a) Database b) Knowledge c) Queries d) Transactions

(vii) Identify the type of learning in which labelled training data is used. **[CO3]**

- a) Semi-supervised Learning
- b) Supervised Learning
- c) Reinforcement Learning
- d) Unsupervised Learning

(viii) Modeling ordered values comes under **[CO3]**

- a) Classification
- b) Prediction
- c) Supervised
- d) Reinforcement

(ix) Euclidean Distance can only be used if attributes are **[CO3]**

- a) Continuous
- b) Finite
- c) Discrete
- d) Infinite

(x) Learning by Observation is also known as **[CO3]**

- a) Classification
- b) Clustering
- c) Regression
- d) Prediction

(xi) Simple Matching Coefficient is used when binary variable is **[CO5]**

- a) Asymmetric
- b) Symmetric
- c) in-variant
- d) non-invariant

(xii) k-medoids comes under which method **[CO5]**

- a) Partitioning
- b) Hierarchical
- c) Temporal
- d) Pattern

(xiii) Based on the precision of the periodicity, a pattern can be divided into **[CO5]**

- a) 2 types
- b) 3 types
- c) 4 types
- d) 5 types

(xiv) Long-Memory Time-Series Modeling is a method for **[CO4]**

- a) Time-Series
- b) Periodicity Analysis
- c) Trend Analysis
- d) Similarity Search

(xv) Under similarity search, dimensionality reduction techniques include **[CO4]**

- a) DAT
- b) DCT
- c) SVD
- d) SVM

(xvi) Full form of FAISS is **[CO4]**

- a) Facebook AI Similarity Search
- b) Facebook Augmented Similarity Search
- c) Facebook AI Similarity Scheme
- d) Facebook AI Similarity Scan

(xvii) Similarity Search finds which type of data **[CO6]**

- a) Exact
- b) Similar
- c) Mixed
- d) None of the above

(xviii) Full form of ARIMA is **[CO4]**

- a) Auto-Regressive Integrated Moving Average
- b) Auto-Regressive Installed Moving Average

- c) Auto-Regressive Integrated Mining Average
- d) All-Regressive Integrated Moving Average

(xix) Which of the following is not the property of stream data? **[CO4]**

- a) Vast Volume
- b) Static Nature
- c) Possibly Infinite
- d) Multidimensional Features

(xx) Which of the following are not the methods for stream data processing & systems? **[CO3]**

- a) Reservoir Sampling
- b) Multi-resolution Models
- c) Sketches
- d) Static Algorithms

(xxi) Bucket width in Lossy Counting Algorithm is **[CO5]**

- a) $\text{ceil}(1/\epsilon)$
- b) $\text{floor}(1/\epsilon)$
- c) $\text{ceil}(1*\epsilon)$
- d) $\text{floor}(1*\epsilon)$

(xxii) In lossy counting algorithm, frequency list is stored in **[CO5]**

- a) Secondary Memory
- b) Cache Memory
- c) Primary Memory
- d) All of the above

(xxiii) Hoeffding tree algorithm runs in **[CO5]**

- a) Linear Time
- b) Exponential Time
- c) Sub linear Time
- d) Can't say about time

(xxiv) What is the full form of VFDT? **[CO6]**

- a) Very Frequent Decision Tree
- b) Volatile Frequent Decision Tree
- c) Very Fast Decision Tree
- d) Very Far Decision Tree

(xxv) The Polysemy Problem is very common in **[CO4]**

- a) Frequent Pattern Mining
- b) Data Warehousing
- c) Graph Mining
- d) Web Mining

(xxvi) In social network analysis, the out-degree and in-degree both follows **[CO5]**

- a) Heavy-tailed Distribution
- b) Light-tailed Distribution
- c) Middle-tailed Distribution
- d) None of the above

(xxvii) Full form of DOM is **[CO6]**

- a) Data Object Model
- b) Document Offer Model
- c) Document Object Model
- d) Document Object Mining

(xxviii) In class imbalance problem the number of positive and negative class is **[CO3]**

- a) Equal
- b) Not equal
- c) Depends on situations
- d) No relation

(xxix) Full form of FSG is [CO4]

- a) Frequent Sub-graph Mining
- b) Fast Sub-graph Mining
- c) Fast Sub-graph Method
- d) Frequency Sub-graph Method

(xxx) What is the full form of VIPS? [CO6]

- a) Vision Based Page Segmentation
- b) Value Based Page Segmentation
- c) Vision Based Page Sequencing
- d) Value Based Program Segmentation

Group – B

(Multiple Choice Type Questions)

2. Choose the correct alternatives from the followings:

10 x 2 = 20

(i) A sequence α is frequent in sequence database S if [CO5]

- a) $\text{Support}_S(\alpha) > \text{min_sup}$
- b) $\text{Support}_S(\alpha) = \text{min_sup}$
- c) $\text{Support}_S(\alpha) \geq \text{min_sup}$
- d) $\text{Support}_S(\alpha) \leq \text{min_sup}$

(ii) Which type of clustering technique is applied in BIRCH? [CO5]

- a) Multi Order
- b) Single Order
- c) Single Phase
- d) Multi Phase

(iii) In trend analysis, Moving Average comes under which of the following characteristics [CO4]

- a) Cyclic Movements
- b) Seasonal Variations
- c) Trend Movements
- d) Random Movements

(iv) In AGM algorithm, two size- k frequent graphs are joined only if they have [CO3]

- a) Same Size- $(k-1)$ Subgraph
- b) Different Size- $(k-1)$ Subgraph
- c) Same Size- $(1-k)$ Subgraph
- d) Optimal Size- $(k+1)$ Subgraph

(v) Compared with DOM-based methods, the segments obtained by VIPS are [CO4]

- a) More Semantically Spreaded.
- b) Less Semantically Aggregated.
- c) More Systematically Aggregated.
- d) More Semantically Aggregated.

(vi) gSpan algorithm is a [CO6]

- a) Apriori-Based Approach
- b) Pattern Growth Approach
- c) Classification
- d) Clustering

(vii) Which is the correct limitation of Hoeffding Tree Algorithm? [CO3]

- a) The algorithm cannot handle concept drift
- b) It takes exponential time
- c) A node can't be splitted more than once
- d) None of the above

(viii) The accuracy of ensemble approach is [CO3]

- a) Greater Than Single Classifier
- b) Lesser Than Single Classifier
- c) Greater Than Multiple Classifier
- d) No Relation

(ix) Drawback of using probabilistic hierarchical clustering is [CO5]

- a) It gives detailed analysis than other clustering approach which may be problematic
- b) Various probability models are used
- c) It cannot handle the uncertainty of cluster hierarchies
- d) All of the above

(x) In K-Means the quality of cluster C_i can be measured by [CO5]

- a) Sum of squared error
- b) Absolute squared error
- c) Squared error
- d) Any one of the above

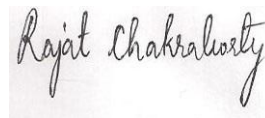
Group – C

(Short Answer Type Questions)

3. Attempt any four from the followings:

4 x 5 = 20

- i) Explain Data Mining Architecture with proper diagram. [CO1]
- ii) (a) Explain Agglomerative Clustering with diagram. (2 Marks) [CO5]
(b) Explain Divisive Clustering with diagram. (3 Marks) [CO5]
- iii) What are the major components of trend analysis? [CO4]
- iv) Explain Lossy Counting Algorithm. [CO3]
- v) What are the major characteristics of Social Network Analysis? [CO6]
- vi) How to mine multimedia data present on the web? [CO4]



Signature of the Paper Setter/Moderator

Answer Key:

Group – A

- | | | | | | |
|-----------------|----------|---------|----------|-----------|-----------|
| 1. (i) a | (ii) d | (iii) d | (iv) d | (v) d | (vi) b |
| (vii) b | (viii) b | (ix) a | (x) b | (xi) b | (xii) a |
| (xiii) a | (xiv) c | (xv) c | (xvi) a | (xvii) b | (xviii) a |
| (xix) b | (xx) d | (xxi) a | (xxii) c | (xxiii) c | (xxiv) c |

(xxv) d	(xxvi) a	(xxvii) c	(xxviii) b	(xxix) a	(xxx) a
---------	----------	-----------	------------	----------	---------

<u>2.</u> (i) c	(ii) d	(iii) c	<u>Group – B</u> (iv) a	(v) d	(vi) b
(vii) a	(viii) a	(ix) c	(x) a		

<u>3.</u> (i)	(ii)	(iii)	<u>Group – C</u> (iv)	(v)	(vi)
---------------	------	-------	--------------------------	-----	------