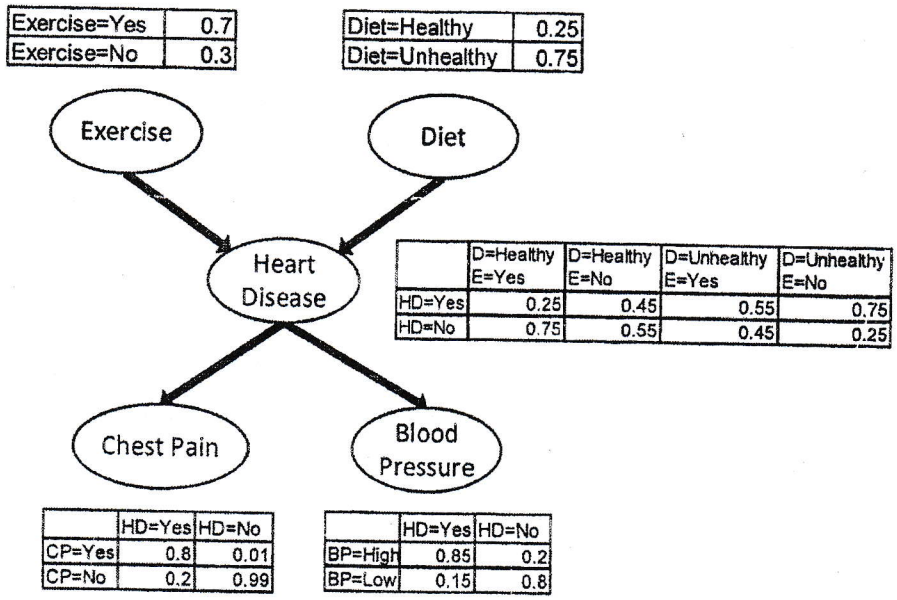
TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2081 Baishakh

Exam.	Back		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining (*CT72502*) (*Elective I*)

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. a) What is Data mining? Explain the steps of KDD process briefly. [2+5]
   b) What is Data Pre-Processing? Briefly explain the major tasks performed in data pre-processing. [2+7]
2. a) How does Neural Net work classified work? Explain with suitable example. [8]
   b) What is limitation of Naive Bayes and how Bayesian Belief Networks overcomes it? If a person does exercise, eats an unhealthy diet and has blood pressure but no chest pain, will that person has a heart disease? [3+5]



3. a) When do we use Association analysis? Explain FP-Tree with an example. [2+5]
   b) What is limitation of Apriori algorithm compared to FP-growth? A database has 5 transactions, given in table below. Let min support = 60% and min confidence = 80%. [2+7]

TID	Item_bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

   i) Find all frequent itemsets using FP-growth.
   ii) List all of the strong association rules (with support and confidence).
4. a) What is Cluster Analysis? What are its applications? Explain different types of clusters. [8]
   b) Use K-means clustering to cluster the following given data for K = 2 with Euclidean distance matrix. List down the demerits of this algorithm. [5+3]
5. a) Explain briefly the key steps in text mining. How do you find page rank? Explain. [4+4]
   b) What is Anomaly Detection? Why is Anomaly Detection important? Briefly explain different types of anomaly detection schemes. [2+2+4]

***

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2080 Bhadra

Exam.	Regular		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data mining (*CT72502*) (*Elective I*)

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. a) Write key features of data warehouse. Explain each steps of knowledge discovery data mining process with a suitable example. [2+5]

   b) How do similarity / dissimilarity is calculated? Find the cosine similarity between Object-2 and 4. Also calculate the Euclidean distance between object 1, 3 and object 1, 4. [3+3+3]

Object	Size	Weight	Color_Code	Taste_Score
1	4	56	7	10
2	3	53	8	11
3	7	58	6	9
4	9	55	7	12

2. a) How does Rule Based Classifier work? Explain with suitable example. [7]

   b) When do we use classifier? You have the following information about the flower. Your job is to classify the given flower with SepalLength = 7, SepalWidth = 3.2, PetalLength = 4.7, and PetalWidth = 1.4. Use KNN algorithm for K = 3 with Euclidean distance matrix. [3+6]

Id	Sepal_Lenth	Spal_Width	Petal_Lenth	Petal_Width	Labal
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	6	2.2	4	1	versicolor
5	6.1	2.9	4.7	1.4	versicolor
6	5.6	2.9	3.6	1.3	versicolor
7	6.7	3.1	4.4	1.4	versicolor

3. a) What is FP-Growth Algorithm? Explain FP-Growth Algorithm with example. [2+5]

   b) What is Association Analysis? Explain with different use cases? Use the Apriori algorithm to find the frequent itemsets. Assume minimum support count is 4 and confidence is 80%. [2+7]

TID	Items
T100	F, A, C, D, G, I, M, P, N
T101	A, B, C, D, F, L, M, O, P
T102	B, F, H, V, J, O, P
T103	B, C, K, S, A, V
T104	L, A, F, C, E, P, M, N, V
T105	I, B, A, P, S, M
T106	F, A, C, I, B, A, P

4. a) When do we use clustering? How do you evaluate the cluster generated? [3+4]

b) What is hierarchical clustering? Use this clustering approach to draw dendrogram for given data points. [2+7]

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

5. a) What is Web Mining? Briefly explain structure of Web Mining. [3+5]

b) Explain different types of outlier with suitable examples. How density based outlier detection works? [5+3]

***

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2080 Baishakh

Exam.	Back		
Level	BE	Full Marks	80
Programme	BCT, BEX	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining (*CT72502*) (*Elective-I*)

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What is data warehousing? Explain with example where data warehouses are used. [2+4]
2. a) List down the different types of similarity measures by highlighting their application areas. [3]

   b) Consider the following table: [2+2+2+1]

Name	Gender	Eyecolor	Haircolor	Test-1	Test-2	Fever	Cough
Ram	M	Black	Gray	P	N	P	N
Laxmi	F	Blue	Black	P	P	N	N
Shyam	M	Blue	Gray	N	P	N	P

   i) Calculated Jaccard Coefficient of Sim_{jacard} (Ram, Laxmi) for asymmetric binary attributes.
   ii) Dissimilarity of symmetric binary attributes d (Laxmi, Shayam).
   iii) Find the Simple matching coefficients of SMC (Ram, Shyam).
   iv) Find the Cosine similarity between documents d1 = (4, 1, 2, 0, 2, 0, 0) and d2 = (2, 1, 3, 0, 1, 1, 1)

3. In what cases you cannot use Accuracy for performance measure, give some examples. Assume that you have the following confusion matrix. Calculate the Classification error, Sensitivity, False alarm rate Specificity. [3+5]

		Actual Values	
		True	False
Predicated Values	True	1050	250
	False	150	950

4. What is Nearest Neighbor Classifier? What are the main issues with this classifier? Propose another classifier that solves the issues. [1+3+4]
5. Generally, we will be more interested in associated rules with high confidence. However, often we will not be interested in association rules that have a confidence of 100%. Why? Then specifically explain why association rules with 99% confidence may be interesting (i.e., what might they indicate)? Identify the candidate and large item sets of the following transaction table using Apriori algorithm with minimum support 2. [4+5]

TID	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

6. Where is association analysis applicable and beneficial for us? Elaborate FP Growth Method Algorithm with examples. [2+6]
7. Cluster the following samples based on complete-linkage algorithm and draw the dendrogram. Using Euclidean distance. [8]

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

8. Describe K-means algorithm for clustering and discuss strategy in determining the optimal value of K. [4+4]

9. What is anomaly detection? Explain distance based method for anomaly detection. [2+3]

10. Write short notes on the following: [2×5]

a) Neural Network Classifier
b) Time Series Data Mining

***

TRI BHUVAN UNIVERSITY
INSTITUT E OF ENGINEERING
Examinati n Control Division
2)79 Bhadra

Exam.	Regular		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining (*Elective I*)(*CT72502*)

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt All questions.*
- ✓ *The figures in the margin indicate Full Marks.*
- ✓ *Assume suitable data if necessary.*

1. What is Data Mining? What are the steps involved in knowledge discovery process? [1+5]
2. Explain typical OLAP operations over a multidimensional data warehouse? Differentiate between OLAP and OLTP tools. [6+4]
3. Draw decision tree for the given data using ID3 algorithm. [10]

Age	Income	Student	Credit_Rating	Buy's_Computer
Youth	High	No	Fair	No
Youth	High	No	Excellent	No
Middle_Aged	High	No	Fair	Yes
Senior	Medium	No	Fair	Yes
Senior	Low	Yes	Fair	Yes
Senior	Low	Yes	Excellent	No
Middle_Aged	Low	Yes	Excellent	Yes
Youth	Medium	No	Fair	No
Youth	Low	Yes	Fair	Yes
Senior	Medium	Yes	Fair	Yes
Youth	Medium	Yes	Excellent	Yes
Middle_Aged	Medium	No	Excellent	Yes
Middle_Aged	High	Yes	Fair	Yes
Senior	Medium	No	Excellent	No

4. Suppose you have a test record "X = (Home Owner = No, Material Status = Married, Income = $120K)". Your job is to classify this record using Naive Bayesian Classification. Use the following table for your calculations. [6]

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

5. Derive association rule for the following market basket transactions. [8]
   Minimum support = 50%
   Minimum confidence = 80%

Transaction ID	Item Set
1	A,B
2	A,D
3	A,C
4	B,E
5	B,D,E
6	A,E,C

6. a) How do you handle the categorical attributes in data mining process? Explain with example. Generate the at least four subsequences from the given sequence: < {2,3,5}, {6,7,8}, {9,1}, {7,4} >. [3+3]

   b) What are subgraph pattern? [2]

7. What are core, border and noise points? Write the algorithm of DBSCAN clustering and explain how it is useful in handling the noisy data. [3+5]

8. An internet marketer is interesting in segmenting internet based the input attributes – top ten search key words used, top 10 URLs, recent 10 online purchases (vendor, product, qty, amt), Internet usage level, heaviest access hour, and heaviest access day of a week. Which clustering algorithm do you think can be used for segmentation? How do you validate the cluster which has been created? [2+6]

9. What do you mean by anomaly detection? Why is it important and where is it applicable? [3+3]

10. Write short notes on: [2×5]

    a) Page Rank algorithm
    b) FP-Tree

***

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2079 Baishakh

Exam.	Back		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. Describe the process of knowledge discovery in databases. Explain the specific challenges that motivated the development of data mining. [3+3]
2. Suppose that the data for analysis include the attribute the frequency of stop words in documents. The values are given in increasing order: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. [6]
   a) Use smoothing by bin means with a depth of 3.
   b) Use min-max normalization to transform the value 35 into the range from 0.0 to 1.0.
   c) Use z-score normalization to transform the value 35 where the standard deviation of the above frequency is 12.94.
   d) Use normalization by decimal scaling to transform the value 35.
3. What are nominal and ordinal attributes? Discuss how to handle missing data and noisy data during data cleaning process. [6]
4. Describe the working mechanism as well as the merits and demerits of the holdout method, random sampling, k-cross validation and bootstrap approaches in evaluating the performance of a classifier. [8]
5. a) Write Apriori algorithm and using the algorithm find all the frequent itemset for the following database. (min_sup = 20%). [5]

TID	A1	A2	A3	A4	A5	A6	A7	AB	A9
T2	0	1	0	1	0	0	0	1	0
T3	0	0	0	1	1	0	1	0	0
T4	0	0	1	0	0	0	0	0	0
T5	0	0	0	0	1	1	1	0	0
T6	0	1	1	1	0	0	0	0	0
T7	0	1	0	0	0	1	1	0	1
T8	0	0	0	0	1	0	0	0	0

   b) Is it possible to generate any rules out of the frequent items, considering any value of confidence threshold? [3]
6. Given the following confusion matrix, determine Accuracy, Error rate, Sensitivity, Specificity, Precision, Recall of the classifier model. [4]

n = 165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

7. How does FP growth approach generate frequent item sets without generating candidate item sets? Explain with an example. [6]

8. Define clustering. Given the matrix whose column represent different data points, perform a K-means clustering on this dataset using the Manhattan as the distance function. The center of the 3 clusters are initiated as A(6.2, 3.2), B(6.6, 3.7) and C(6.5, 3.0). Provide the final cluster centers after 3 iteration. [8]

5.9	4.6	6.2	4.7	5.5	5.0	4.9	6.7	5.1	6.0
3.2	2.9	2.8	3.2	4.2	3.0	3.1	3.1	3.8	3.0

9. Explain hierarchical clustering method with an example of Dendogram plot. [6]

10. Describe the strengths and weaknesses of the statistical, proximity-based, density-based and cluster-based approaches of anomaly detection. [6]

11. Write short notes on: [4×4]

a) Time Series data mining
b) Web mining
c) Data visualization
d) DBSCAN clustering

***

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2078 Bhadra

Exam.	Regular		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I)(CT 72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. Explain how data mining system can be integrated with database/data warehouse system. Explain Data mining process with diagram. [4+2]

2. Suppose that a data warehouse consists of the four dimensions data, spectator, location, and game, and the two measures count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults or seniors, with each category having its own charge rate. [3+3]

   a) Draw a star schema diagram for the data warehouse.
   b) Starting with the base cuboid [data, spectator, location, game], what specific OLAP operations should you perform in order to list the total charge paid by student spectators at Dashrath Stadium in 2021?

3. Use the following methods to normalize the data: 200, 300, 400, 600 and 1000. [2+2+2]

   a) Min-max normalization by setting min=0 and max=1
   b) Z-score normalization
   c) Normalization by decimal scaling

4. Construct a decision tree for the following data set using information gain. [8]

   Predict the class label for a data point with values<Female, 2, standard, high>

Gender	Car ownership	Travel cost	Income level	Transport mode
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car
Female	1	Cheap	Medium	Train
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train

5. Consider the given transactional database from a grocery store. Use a support threshold of 33.34% and confidence threshold of 60% to compute the following: [4+4]

   a) Build a frequent pattern tree (FP-Tree). Show for each transaction how the tree evolves.
   b) Use FP-Growth algorithm to discover the frequent itemsets from this FP-tree.

Transcation ID	Items
T1	HotDogs, Buns, Ketchup
T2	HotDogs, Buns
T3	HotDogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	HotDogs, Coke, Chips

6. Calculate: Accuracy, TPR, FPR and Precision for the given confusion matrix for a classifier. [4]

		Actual Class	
		Class 1	Class 2
Predicted Class	Class 1	142	40
	Class 2	98	720

7. Explain Naive Baiyesian classification algorithm with suitable example. [6]

8. Write K-means clustering algorithm. Generate two clusters from following dataset using K-means clustering. [2+6]

Instance	A	B
1	1	2
2	2.5	1
3	3.5	1.5
4	4	1
5	3.5	2.5
6	5	3

9. Provide answers to the following with regard to the DBSCAN clustering approach: [2+2+2]
   a) How does the DBSCAN quantify the neighborhood of an object? How is a large dense region assembled from small dense regions centered by core objects?
   b) How does DBSCAN find clusters? How are the neighborhood threshold (Epsilion) and minimum number of points (MinPts) determined empirically in DBSCAN?
   c) Prove that in DBSCAN, for a fixed minimum number of points (MinPts) value and two neighborhood thresholds, Epsilion1 < Epsilion2, a cluster (C) with respect to Epsilion1 and MinPts must be a subset of a subset of a cluster (K) with respect to Epsilion2 and MinPts.

10. Compare and contrast among three difference methods of anomaly detection. [6]

11. Write short notes on: [4×4]
    a) Minkowski Distance
    b) Laplacian Correction in Classification method
    c) Page rank algorithm in Web mining
    d) Overfitting problem in classification

***

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2076 Ashwin

Exam.	Back		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What are the fundamental differences between Data Mining and Data Warehousing? Describe the steps of KDD for data mining. [3+7]
2. What do you mean by dimensional data? What are base & apex cuboid? Slicing & Dicing? Roll Down and Roll UP operations? Give example. [2+3+3+3]
3. How do you measure the accuracy of classifiers? How do you select best root attribute in decision tree? Explain. [4+6]
4. What are prior and posterior probabilities? Explain the algorithmic steps of Bayesian classifier and write its strengths. [3+7]
5. For the transactions given below, consider confidence=60% and minimum support=30%. Identify large itemsets (L-Itemset) at L=3 with possible associations using A-priori algorithm and generate F-List using FP-Growth algorithm. [12]

Transactions	Items description
T1	A, B, C, T, M, P, D, K
T2	A, B, T, P, D, K
T3	B, C, T, D, M, A, P
T4	A, C, T, M, D,
T5	A,C, D, K, M
T6	B, C, T

6. How DBSCAN algorithm works? How do we avoid the issues of DBSCAN? [8+2]
7. Explain web mining taxonomy. [8]
8. Write short notes on (**Any Three**) [3+3+3]
   a. Data smoothing techniques
   b. Clustering and its application in anomaly detection
   c. AprioriALL: Sequential pattern mining algorithm

(30)

26B TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2075 Ashwin

Exam.	Back		
Level	BE	Full Marks	80
Programme	BCT, BEX	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. How is data warehouse different from a database? How are they similar? [2+2]
2. Discuss issues to consider during Data Integration. Describe OLAP and operations on OLAP with suitable example. [5+5]
3. Explain Naïve Bayesian classification with suitable example. [8]
4. The confusion matrix for a classifier is given as follows: [10]

		Predicted Class	
		Class 1	Class 2
Actual Class	Class 1	21	6
	Class 2	7	41

Calculate: Accuracy, Sensitivity, Specificity and Precision.

5. Why association analysis is required in data mining? Explain Apriori principle with example. [2+6]
6. What are the advantages of FP growth method? Explain FP growth algorithm. [2+6]
7. Explain K-means clustering with limitation. Generate two clusters from following dataset using K-means clustering. [4+6]

A	B
1	2
2.5	4.5
4	6
3.5	4
4	5.5
3	6

8. What are outliers? Explain an algorithm that can be used to generate density based clusters. [8]
9. Why anamoly detection is important? Explain distance based method for anamoly detection. [2+6]
10. Explain Web mining and Multimedia mining. [6]

***

(30)

36 B TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2074 Ashwin

Exam.	Back		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What is data mining? Explain the process of data mining. [2+3]
2. In real-world data, tuples with missing values for same attributes are a common occurrence. Describe various methods for handling this problem. [5]
3. What is classification? Explain Rule-Based classification with its classification principles with suitable example. [2+8]
4. The confusion matrix for a classifier is given as follows: [10]

		Predicted Class	
		Class 1	Class 2
Actual Class	Class 1	25	9
	Class 2	4	31

Calculate:

a) Accuracy
b) Sensitivity
c) Specificity
d) Precision

5. Identify the candidate, frequent item sets and association rules for the following transaction data using Apriori algorithm. [8]

TID	ITEMS
1	M1, M2, M5
2	M2, M4
3	M2, M3
4	M1, M2, M4
5	M1, M3
6	M2, M3
7	M1, M3
8	M1, M2, M3, M5
9	M1, M2, M3

Take minimum support = 20%, minimum confidence 80%

6. Explain FP-Growth algorithm with example. [8]

7. Write K-means algorithm and find clusters for following data set. [2+8]

Instance	X	Y
1	1.0	2.0
2	2.5	1.0
3	3.5	1.5
4	4.0	1.0
5	3.5	2.5
6	5.0	3.0

(Take K = 2)

8. What is web mining? Explain different categories of web mining. [6]

9. List the various types of partition based clustering methods. Explain Hierarchical clustering method with an example. [10]

10. Write short notes on: (Any two) [2×4]

a) OLAP Operations
b) Density reachable and Density Connected
c) Data Mining for Anomy Detection

***

(30)

26B TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2074 Chaitra

Exam.	Regular		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What is data warehouse and data mart? Describe Snowflake scheme with example. [2+4]
2. What are the approaches to handle missing data? Describe OLAP and operations on OLAP with suitable example. Differentiate between OLAP and OLTP. [2+5+3]
3. Draw clear block diagram depicting different stages in classification. Explain the inverse relation between precision and recall. Given the confusion matrix, determine accuracy, sensitivity and precision of the classifier model. [2+3+5]

Actual \ Predicted	Positive	Negative
Positve	142	40
Negative	98	720

4. Explain decision tree with the concept of Naive base classification with appropriate example. [10]
5. Why association analysis is required in data mining? Explain apriori principle with example. [2+6]
6. How does FP growth approach overcomes the disadvantages of Apriori algorithm. For the transaction data given in table generate FP-Tree. [2+8]

Transaction ID	Item set
T1	Camera, Laptop, Pen drive
T2	Laptop, Pen drive
T3	Laptop, Mobile, Earphone
T4	Earphone, Mobile
T5	Camera, Earphone
T6	Laptop, Mobile, Earphone

7. Describe the difference between Hierarchical and partitioning clustering. How K-means clustering is applied? Verify using example. [2+8]
8. What do you mean by anomaly detection and why is it important? Describe distance based approaches for anomaly detection. [4+3]
9. Write short notes on: (any three) [3×3]
   i) Issues in clustering
   ii) Multimedia mining
   iii) Time series data mining
   iv) Web mining

***

37B TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2073 Shrawan

Exam.	New Back (2066 & Later Batch)		
Level	BE	Full Marks	80
Programme	BE, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective II) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. "The world is data rich but information is poor". Justify with your own words. [8]
2. What are the measuring elements of data Quality? Explain different data transformation by normalization methods with an example. [2+6]
3. What is a decision tree and how information gain is used for attribute selection? Explain with example. [8]
4. Explain ROC. Using the following data, calculate TPR, FPR, precision for given confusion matrix. [1+3+6]

	A	B
A	20	5
B	10	40

Classify, A = Yes, B = No

5. What is FP Tree? How FP-growth algorithm eliminate the problem of Apriori algorithm? Construct the FP tree and find association rules for the following transaction database using FG- Growth algorithm. Support = 30% and confidence = 75%. [10]

Transaction ID	Items
1	P,R,S
2	R,S,T
3	P,Q,R
4	P,R,S,T
5	P,S,T
6	P,Q,T
7	Q,S,T
8	Q,R,T

6. What are Categorical data? What are the possible issues arrives when using Categorical data? How can you handle such issues? [2+3+3]
7. What is the application of clustering in data mining? Explain the k-means algorithm with example. [8]
8. What is anamoly detection? Explain distance based method for anamoly detection. [8]
9. Write short notes on: [4×3]
   i) Data transformation
   ii) Web mining
   iii) OLAP

***

(30)

36 B TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2073 Chaitra

Exam.	Regular		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. How is data warehouse different from RDBMS? Also list the similarities. [2+2]
2. What is data pre-processing? Explain data sampling and dimensionality reduction in data pre-processing with suitable example. [2+4+4]
3. How data in most real application becomes Asymmetric. Explain the difference between symmetric and asymmetric data. [5]
4. What is ID3 algorithm? Calculate TPR, FPR and Accuracy for given confusion matrix. [2+6]

	Predicted +	Predicted -
Predicted +	100	40
Predicted -	60	300

5. Explain Apriori algorithm in market basket analysis? Derive association rule from the following market basket transactions with 50% of minimum support and confidence respectively. [3+7]

Transaction	Itemsets
1	A, B, C
2	A, C
3	A, D
4	B, E, F

6. What is the use of FP-Growth method in market basket analysis? Explain FP-Growth method with a suitable example. [10]
7. How clustering differ from classification? Given the one-dimensional points {5, 12, 18, 24, 30, 42, 48} with initial centroids {5, 12, 18}, create three clusters by K-Means algorithm and calculate SSE for this clustering result. [4+8]
8. Explain Sequential Pattern and Sub-graph Pattern with suitable example. [4+4]
9. What is anomaly detection? Explain the issues associated with anomaly detection. [2+3]
10. Write short notes on: (Any two) [2×4]
    a) Time series data mining
    b) Overfitting and ROC
    c) www mining

***

27C TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2072 Kartik

Exam.	New Back (2066 & Later Batch)		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What is a data mining? Explain general steps in brief. [4]
2. Why data preprocessing is required in the data mining? Explain some of approaches of data clearing. [5+5]
3. Write about Hunt's Algorithm for Decision Tree induction. Explain the test conditions that can be used for different attribute types. [10]
4. What is an ANN classifier? Explain its general consideration that required for the classifier. [2+6]
5. What is an association analysis? Explain its importance in market-basket analysis. [2+5]
6. What is a Frequent item set? Explain FP growth method with example. [1+8]
7. What is a cluster analysis? How it is different from classification? [5]
8. Explain a DBSCAN algorithm with example. [7]
9. What is an Anomaly detection? Discuss its importance in security. [5]
10. Explain Time series data mining in brief. [6]
11. Write short notes on: [3×3]
    i) Data transformation
    ii) Sequential pattern
    iii) Cluster evaluation

***

(30)

34B TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2072 Chaitra

Exam.	Regular		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective II) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What is data mining? Explain all the steps of knowledge discovery. [2+6]
2. How do you perform analysis of multidimensional data? Explain with the concept of OLAP. [10]
3. Predict Class label using naive Bayesian classifier for X = (age = youth, income = medium, student = yes, credit-rating = fair) using the following data set. [10]

RID	Age	Income	Student	Credit-rating	Class Buy computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle-age	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle-age	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle-age	Medium	No	Excellent	Yes
13	Middle-age	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

4. The confusion matrix for a classifier is given as follows: [10]

		actual class	
		class1	class2
predicted class	class1	21	6
	class2	7	41

calculate a. accuracy
b. sensitivity
c. specificity
d. precision
e. recall

5. What is the importance of SUPPORT and COFIDENCE during association analysis? Explain FP-Growth method with example. [10]
6. What are the types of clustering methods? Explain DBSCAN method of clustering with an example. [10]
7. What is the use of Apriori Algorithm in market basket analysis? Explain with suitable example. [10]
8. Write short notes on: [4×3]

   i) Time series Data mining
   ii) Issues in anomaly/Fraud detection
   iii) Categorical data and related issues

***

20

27C TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2071 Shawan

Exam.	New Back (2066 & Later Batch)		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(CT72502) (Elective I)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ ***All** questions carry equal marks.*
- ✓ *Assume suitable data if necessary.*

1. What is data mining? Explain different data types of attributes in a dataset.
2. How can principle component analysis be used for dimentionality reduction?
3. Why is classification a super vised learning method? Explain different impurity measures used in decision tree classifier.
4. Explain Naive Bayes classifier. How can over fitting problem be solved in case of classification?
5. Explain FP-growth algorithm in detail.
6. What are association rules? How can spriori algorithm be used to generate association rules.
7. What is contiguous cluster? Explain an algorithm that can be used to generate contiguous clusters.
8. Explain K-means clustering with limitation Use k-means clustering to cluster the following dataset.

A	B
1.0	1.0
1.5	2.0
3.0	4.0
5.0	7.0
3.5	5.0
4.5	5.0
3.5	4.5

9. How can Nearest-Neighbor algorithm be used for anomaly defection?
10. Write short notes on:
    a) Time-series data mining
    b) Data warehouse and data mart

***

20

24C TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2071 Chaitra

Exam.	Regular		
Level	BE	Full Marks	80
Programme	BEX / BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ ***All** questions carry equal marks.*
- ✓ *Assume suitable data if necessary.*

1. What is a Data Mining? Explain its application.
2. Explain the properties that a Distance Metric needs to support with respect to Minkowski's distance.
3. What is a decision tree? Explain Gini Index with suitable example.
4. Explain a Bayes classifier. In what cases can Naive Bayes and Bayesian Belief Network be used?
5. Why is a clustering an unsupervised learning? How can hierarchical clusters be generated using Bisecting K-means algoritlun?
6. Explain the different measures of cluster validity.
7. How does Apriori Algorithm optimize the brute force approach for frequent item set generation?
8. What is an Anomaly Detection? Explain few distance based approaches that can be used for Anomaly Detection.

***

(20)

02C TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2070 Ashad

Exam.	New Back (2066 & Later Batch)		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt All questions.*
- ✓ *The figures in the margin indicate Full Marks.*
- ✓ *Assume suitable data if necessary.*

1. What is dimensionality reduction? Why is it important in data mining? [5]
2. What is the importance of homogeneousness measure in decision tree classifier? Explain GINI index? [8]
3. What are the properties of a distance metric? How is distance metric used in instance based classifier? [10]
4. What are association rules? Explain its importance with example. [6]
5. Explain FP growth algorithm in detail with example. [12]
6. What are density based clusters? Explain DBSCAN clustering algorithm with example. [10]
7. Explain different measures that can be used to compare two clusters. [6]
8. What is anomaly desertion? Explain likelihood approach for anomaly detection. [8]
9. Explain seasonality in time series data. [5]
10. Write short notes on: [5+5]
    a) OLAP cubes
    b) Data ware house

***

(20)

INSTITUTE OF ENGINEERING

Examination Control Division

2070 Chaitra

Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year/Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT725)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt <u>All</u> questions.*
- ✓ *The figures in the margin indicate <u>Full Marks</u>.*
- ✓ *Assume suitable data if necessary.*

1. a) What is "curve of Dimensionality"? How can it be avoided? [5]

   b) Discuss the impact of noisy data in data mining? [5]

2. Explain rule based classifier? How can CN2 Algorithm be used for rule based classification? Define "Accuracy" and "Laplace" measures used for rule evaluation. [9]

3. An input sequence "A A B B B A A A B B" was used for classification. The Classifier 'X' predicted the sequences as: "A A B B B A A A B B" where as the Classifier 'Y'predicted the sequences as: "A A A A B B A A A B". Develop the corresponding confusion matrix for the classifiers and find their corresponding. [10]

   i) Accuracy
   ii) Precision
   iii) True Positive Rate
   iv) False Positive Rate

4. Explain Apriori algorithm. Use Apriori to generate frequent item sets with support of 50% for the following transaction database. [10]

TID	Items
1	ACD
2	BD
3	ABCE
4	BDF

5. Why is pattern evaluation important in association rule mining? Explain with example the statistical based measures used for measuring interestingness of association rules. [8]

6. What is a density based cluster. Explain an algorithm that can be used to generate density based clusters. [8]

7. What is Hierarchical Clustering? Differentiate between agglomerative and divisive approach of hierarchical clustering. Augment your answer with appropriate illustrative examples. [10]

8. Write short notes on: [15]

   i) Data ware house and Data mart
   ii) Base Rate Fallacy
   iii) Web mining
   iv) Anomaly Detection
   v) Convex Hull Method

***

37B TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2069 Chaitra

Exam.	Regular		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What are the different data types? Explain with examples. [5]
2. How is decision tree classifier different than rule based classifier? [8]
3. Explain Baye's Theorem. How can it be used for classification? Explain how Naive Baye's simplifier the computational complexity of Baye's classification algorithm. [12]
4. What is frequent item set mining? How do Apriori and FP-growth algorithm optimize the brute force approach for finding frequent item sets? [15]
5. Explain K-means clustering algorithm with examples. [10]
6. Explain the issues regarding cluster validation. [6]
7. What is Base Rate Fallacy? Explain with example. [7]
8. How can Apriori Algorithm be used for finding association rules out of a frequent item set? [7]
9. Write short notes on: [5+5]
   a) Page Rank
   b) Data mart

***