Name of the Course : **B.Sc. (Hons.) Computer Science**

Semester : **VI**

Name of the paper : **Data Mining**

Unique Paper Code : **32347611**

Year of Admission : **2015, 2016, 2017 & 2018**

Duration: 3 Hours **Maximum Marks: 75**

Instructions for Candidates

1. Attempt any FOUR out of SIX questions.
2. All questions carry *equal* marks.
3. Upload single PDF file for each question.

1. Consider a sample dataset of patients visiting a clinic for consultation:

Patient ID	Patient Name	Blood Pressure	Blood Pressure date	Chest Pain	Age	Exercise	Heart Condition
101	Sarita	120	14-10-1995	1	20	High	Good
102	Maddy	110	16-1-2018	0	40	High	Good
103	Rohit	140	18-10-2018	3	?	Low	Bad
104	Gauhar	172	4-6-2018	3	39	Medium	Bad
105	Himani	150	8-6-2018	2	35	Low	Bad
106	Shubham	110	4-7-2018	1	40	Medium	Good
107	Suresh	120	5-3-2018	0	26	High	Good
108	Anmol	110	5-5-2018	1	27	Medium	Good

- Identify the type of each attribute and give justification.
- Perform min-max normalisation on "Blood Pressure" attribute.
- Identify the data quality issues for each of the following fields. Can these issues be resolved? If yes, how?
    - Blood Pressure date for patient 101.
    - The blood pressure meter is miscalibrated and adds 1 mmHg to each reading.
    - Age of the patient 103.

2. A binary classification problem with class labels: 'Yes' and 'No' that denotes the access to the elevator, has the following set of attributes and attribute values:

   Status = {Faculty, Student}
   Floor = {First, Second, Third, Fourth, Fifth}
   Health = {healthy, unhealthy}

   Consider the following set of records for the above classification problem:

Transaction Number	**Status**	**Floor**	**Health**	**Accessible**
1	Faculty	First	Healthy	Yes
2	Student	First	Healthy	No
3	Student	Third	Healthy	No
4	Faculty	Second	Unhealthy	Yes
5	Student	Fifth	Healthy	Yes
6	Faculty	First	Healthy	No
7	Student	Fifth	Unhealthy	No

   A rule-based classifier produces the following rule set:

   R1: Status = Faculty, Floor = Second → Accessible = Yes
   R2: Status = Student, Floor = Second → Accessible = No
   R3: Floor = First → Accessible = No
   R4: Health = unhealthy → Accessible = Yes
   R5: Status = Student, Floor = Fifth, Health = healthy → Accessible = Yes

   - Are the rules in the above rule set mutually exclusive? Justify.
   - Is the rule set exhaustive? Justify.
   - Is ordering needed for this set of rules? Justify
   - Do you need a default class for the rule set? Justify
   - Compute the coverage and accuracy of rules R1 and R5. Which one do you think is a better rule? Why?

3. Consider the training examples shown in the following table for a classification problem.

Student ID	Admission Category	Admission List	Gender	Predicted Class	Actual Class
1	Sports	First	M	C0	C0
2	Arts	Second	F	C0	C1
3	Arts	Third	F	C0	C0
4	Sports	Fourth	M	C0	C1
5	Academics	First	F	C0	C0
6	Arts	Second	F	C1	C1
7	Arts	Third	M	C1	C1
8	Sports	Fourth	M	C1	C0
9	Sports	Third	F	C1	C1
10	Arts	First	F	C1	C0
11	Sports	Third	M	C1	C1
12	Sports	Second	F	C1	C1
13	Sports	First	M	C1	C0

- Compute the Information Gain for the Student ID attribute.
- Compute the Gini Index for the Admission Category Attribute and Admission List attribute
- Which is a better attribute for split based on the Gini Index: Admission Category or Admission List? Why?
- Create a confusion matrix for the above data set and compute False positive rate, accuracy, recall and precision.

4. Consider the following set of points:

{44, 28, 48, 26, 32, 14, 52, 50}

Assuming that k=2, and initial cluster centres for k-means clustering are 5 and 38, compute the sum of squared errors (SSE) and cluster assignment for each iteration.

5. Consider the dataset given below:

Age	Income	Employed	Credit-rating	Buys Car
young	high	yes	fair	yes
young	high	no	good	no
middle	high	no	fair	yes
old	medium	no	fair	yes
old	medium	no	fair	yes
old	low	yes	good	no
old	medium	no	good	no
middle	high	yes	fair	yes

Compute all class conditional and class prior probabilities. Use Naïve Bayes classifier to predict the class of the following tuple:

X = (age = young, income = medium, employed = yes, credit rating = good)

6. Consider the market basket transactions shown in the following table. Use Apriori algorithm to answer the questions that follow.

TID	Item bought
1	oregano, chocolate, milk, cheese, french fries
2	milk, french fries, cheese, ketchup
3	chocolate, cheese, oregano, ketchup
4	chocolate, cheese, french fries
5	french fries, cheese, oregano, chocolate
6	chocolate, ketchup
7	oregano, french fries, ketchup
8	oregano, french fries, chocolate
9	ketchup, oregano, milk
10	french fries, chocolate

- Assuming the minimum support threshold is fixed at 40%, list the set of frequent 1-itemsets (L_1) and with their respective supports.
- List the itemsets in the set of candidate 2-itemsets (C_2) and calculate their supports.
- Generate all association rules from the itemsets in L_2 and also compute the confidence of these rules.

[This question paper contains 12 printed pages.]

Your Roll No. 2202557008S

Sr. No. of Question Paper : 4192 **H**

Unique Paper Code : 2343012005

Name of the Paper : Data Mining – I

Name of the Course : **B.Sc. (Hons.) Computer Science**

Semester : IV

Duration : 3 Hours Maximum Marks : 90

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. **Section A** (Question No. **1**) is compulsory.
3. Attempt any **four** questions from **Section B** (Questions **2** to **7**).
4. The use of a simple calculator is allowed.
5. Parts of the question must be answered together.

P.T.O.

Section A

1. (a) Differentiate between the unsupervised and supervised evaluation measures used for cluster validity. (3)

(b) What is the anti-monotone property of the support measure in association rule mining? Does the confidence measure follow anti-monotone property? (3)

(c) Consider a dataset with two class labels, News and Entertainment, and six labeled documents D1-D6. A new document, D7, is to be classified. The similarity values of D7 with D1, D2, D3, D4, D5 and D6 are 0.75, 0.85, 0.66, 0.87, 0.70 and 0.84 respectively. Using the k-Nearest Neighbor classifier, predict the class label that should be assigned to D7 when k=3. Will the predicted class label change with k=5? (4)

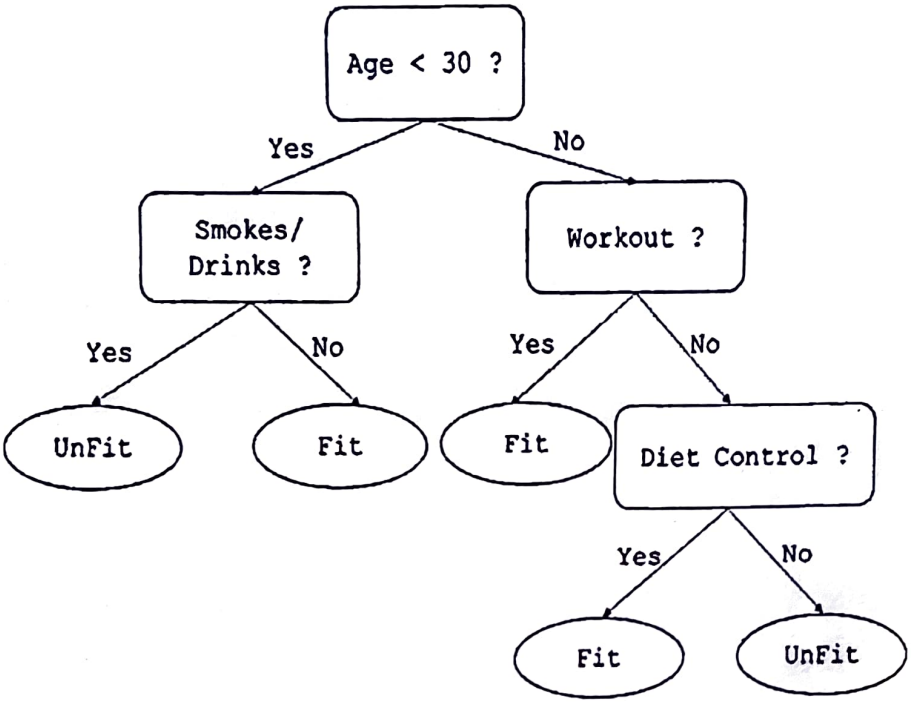
Document	Class Label
D1	News
D2	Entertainment
D3	Entertainment
D4	News
D5	News
D6	Entertainment

(d) Consider the given dataset, which contains six objects, each with two attributes: Age and Salary. K-means clustering is used to cluster the given objects. Do you see any issue with applying K-means to the given dataset? If yes, then state the issue. Also apply the appropriate preprocessing technique to overcome it. If no, state explicitly that no preprocessing technique is required. (4)

	Age (in years)	Salary (in rupees)
Object 1	40	62000
Object 2	24	48000
Object 3	30	54000
Object 4	35	67000
Object 5	46	80000
Object 6	34	66000

(e) Define the curse of dimensionality. The Iris flower dataset comprises of 150 data points and four features, namely sepal length, sepal width, petal width, and petal length. Is it a high-dimensional data or low-dimensional data? Justify your answer. (4)

(f) Consider a decision tree to classify the health of an individual as Fit or Unfit given below :



(i) Extract all classification rules from the decision tree.

(ii) Classify the following object:

Age = 50, Workout = No, Smokes/ Drinks = No, Diet Control = No, Health = ? (4)

(g) Classify the following tasks as "predictive" or "descriptive". Justify your answer. (4)

(i) Foretelling whether an online user will shop on Flipkart for a specific item.

(ii) Grouping the customers of a company according to their buying interests.

(iii) Finding a group of genes such that genes in each group have related functionality.

(iv) Using historical data from previous financial statements to project sales, revenue, and expenses for a company.

(h) Given two objects X = (22, 1, 42, 10) and Y = (20, 0, 36, 8), compute the distance between these two objects using the following distance measures :

(i) Euclidean Distance

(ii) Manhattan Distance (4)

Section B

2. (a) Given the following training dataset, compute all class conditional and prior probabilities. Use the Naive Bayes approach to predict the class label (Salary) for the test instance : (12)

Education Level = PG, Career = Management, Years of Experience = 3 to 10

Education Level	Career	Years of Experience	Salary
UG	Management	Less than 3	Low
UG	Management	3 to 10	Low
PG	Management	Less than 3	High
PG	Service	More than 10	Low
UG	Service	3 to 10	Low
PG	Service	3 to 10	High
PG	Management	More than 10	High
PG	Service	Less than 3	Low
UG	Management	More than 10	High
UG	Service	More than 10	Low

$P(Y=1|X) = P(X|Y) \times P(Y)$

(b) A data mining application uses a particular type of data. Give one application for each of the following type : (3)

(i) Sparse dataset

(ii) Spatio-Temporal data

(iii) Graph-based data

3. (a) Consider the following dataset having details about different departments of a company :

ID	Dept. Name	Location	Established On	Size	Annual Budget
DP12	Finance	Nehru Place	5-01-2020	Large	460
DP19	Marketing	Nehru Place	8-08-2020	Medium	300
DP21	Human Resource	Hauz Khas	2-01-2020	Medium	240
DP27	Production		2-02-2020	Medium	290
DP33	Research & Development	Nehru Place	4-07-2021	Small	90
DP39	Information Technology	Hauz Khas	6-08-2020	Medium	210
DP41	Sales	Nehru Place	9-09-2020	Large	510
DP52	Customer Service	Hauz Khas	2-10-2020	Medium	
DP55	Public Relations	Nehru Place	3-03-2021	Large	900

* Annual Budget is In Lakhs

(i) Identify the type of attributes ID, Dept. Name, Location, Established On, Size, and Annual Budget as nominal, ordinal, interval, or ratio. Give justification for each. (6)

(ii) Suggest a technique for dealing with missing values in the attribute Location. Will the same technique apply to the attribute Annual Budget? Justify. (3)

(iii) What is an outlier? Spot an outlier in the provided dataset. (3)

(b) What is the need for sampling in data mining ? What problems arise if the sample size is too small or too large? (3)

4. Consider the following transactional data of a grocery store :

Transaction ID	Items
T1	Boots, Hoodie, Gloves
T2	Boots, Hoodie
T3	Hoodie, Coat, Cardigan
T4	Cardigan, Coat
T5	Cardigan, Gloves
T6	Hoodie, Coat, Cardigan

(a) What is the maximum number of rules that can be extracted from this data (including rules that have zero support). $3^d - 2^{d+1} + 1$ (3)

(b) Use the Apriori algorithm on the given transactional dataset and compute the candidate and frequent itemsets for each dataset scan. Assume a support threshold of 33.34%. (6)

(c) Enumerate all association rules generated from the largest frequent itemset found in each dataset scan. Compute the confidence of each generated rule. Assuming that the minimum confidence threshold is 70%, find all the strong association rules. (6)

5. (a) A medical team develops classification models for predicting the occurrence of a "genetic disorder" using Classifier A and Classifier B. Patients having genetic disorders are considered positive instances. In contrast, negative instances are ones with the absence of genetic disorders. The classifiers were tested on data from 500 patients and then obtained the result as :

	Actual Label	
	Presence of Genetic Disorder	Absence of Genetic Disorder
Classifier A, predicted "presence of genetic disorder"	131	155
Classifier A, predicted "absence of genetic disorder"	19	195
Classifier B, predicted "presence of genetic disorder"	82	72
Classifier B, predicted "absence of genetic disorder"	68	278

(i) List the confusion matrix for "Classifier A" and "Classifier B". Find the accuracy, precision, sensitivity, recall and specificity for each classifier. (8)

(ii) What problem may occur if the provided training dataset of 500 patients had only 15 positive instances and the remaining negative instances? Which performance measure would you choose to evaluate the classifiers in such a scenario? Which is the better classifier between Classifier A and Classifier B in such a scenario? (4)

(b) Consider a categorical attribute Grade with three values {A, B, and C}. Convert this attribute to asymmetric binary attributes. (3)

6. Consider the given COVID-19 dataset of ten patients.

ID	Age	Fever	BD	Outcome
P1	Young	Yes	High	In ICU
P2	Young	No	High	Hospitalized
P3	Elderly	Yes	High	In ICU
P4	Middle aged	Yes	Moderate	In ICU
P5	Middle aged	No	High	Home Care
P6	Middle aged	Yes	Moderate	In ICU
P7	Elderly	No	Moderate	In ICU
P8	Elderly	No	High	Deceased
P9	Elderly	Yes	High	In ICU
P10	Young	No	High	Hospitalized

BD: Breathing Difficulty

(a) Compute the Gini Index of Age, Fever, and BD attributes. Given that you construct a decision tree using the Gini Index as the splitting criteria, which of the three attributes would you choose at the root? Justify your choice. (9)

(b) Compute the Gini Index of ID. Why should it not be used as a splitting attribute for constructing a decision tree? (3)

(c) Given ten objects in the dataset (P1 – P10), mention all train and test distributions for performing k-fold cross-validation. Assume the value of k = 5. (3)

P.T.O.

7. Given a dataset with six records about startup companies, each record has two fields: Number of Clients and Annual Turnover. Assuming that $k = 2$ and initial cluster centres as the first two records, compute the cluster centres of the resulting clusters until the stopping criterion is met. Use Euclidean distance as the distance metric. Also, compute the SSE (Sum of Squared Error) of each generated cluster.

Number of Clients	Annual Turnover (in Lakhs)
185	72
170	56
168	60
179	68
182	72
188	77

(15)

This question paper contains 7 printed pages]

Roll No. | | | | | | | | | | |

S. No. of Question Paper : **2780**

Unique Paper Code : **32347611** **IC**

Name of the Paper : **Data Mining**

Name of the Course : **B.Sc. (H) Computer Science : DSE-4**

Semester : **VI**

Duration : **3 Hours** Maximum Marks : **75**

(Write your Roll No. on the top immediately on receipt of this question paper.)

Attempt *All* questions from Section A.

Attempt any *four* questions from Section B.

Section A

1. (*a*) Find the Euclidean distance between data points X(0, –1, 0, 1) and Y(1, 0, –1, 0). 2

(*b*) If recall and precision are 0.5 and 0.6 respectively, compute the value of F_1 measure. 2

(*c*) In a given dataset, it is found that an itemset {*ab*} is infrequent. Will itemset {*abc*} be infrequent or frequent ? Explain why. 2

P.T.O.

(*d*) What are the three strategies for handling missing values in a dataset ? 3

(*e*) Differentiate between precision and bias on the basis of the quality of the measurement process. 3

(*f*) What is meant by variable transformation ? What are its advantages ? 3

(*g*) If support of an association rule X -> Y is 80% and confidence is 75%, can we derive support and confidence of the rule Y -> X ? If yes, list down the values. If no, state the reason. 3

(*h*) List down *two* advantages and *two* disadvantages of leave-one-out approach used in cross-validation for evaluating the performance of the classifier ? 4

(*i*) Differentiate between agglomerative and divisive methods of hierarchical clustering with the help of a diagram. 4

(*j*) What are asymmetric attributes ? Give an example of each : 4

(*i*) asymmetric binary attribute,

(*ii*) asymmetric discrete attribute,

(*iii*) asymmetric continuous attribute.

(*k*) The confusion matrix for a 2-class problem is given below : 5

		Predicted Class	
		Class=1	Class=0
Actual Class	Class=1	400	100
	Class=0	200	300

Calculate the Accuracy, Sensitivity, Specificity, True Positive Rate, and False Positive rate.

Section B

2. (*a*) What are the differences between noise and outliers ? Are noise objects always outliers ? Are outliers always noise objects ? 2+1+1

(*b*) Let A and B be two sets of integers. A distance measure '*d*' is defined as follows : 4

$d(A - B)$ = size $(A - B)$ + size $(B - A)$ where '–' denotes set difference. Size denotes the number of elements in the set.

Prove that the distance measure '*d*' is a metric.

(*c*) What is unsupervised learning ? Explain with the help of an example application. 2

3. (*a*) Consider the following dataset for a 2-class problem : 7

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	–
T	T	+
F	F	–
F	F	–
F	F	–
T	T	–
T	F	–

(*i*) Calculate the gain in the Gini Index when splitting on A and B.

(*ii*) Which attribute would the decision tree induction algorithm choose ?

(*iii*) Draw the decision tree after splitting showing the number of instances of each class.

(*iv*) How many instances are misclassified by the resulting decision tree ?

(*b*) Why is K-nearest neighbor classifier a lazy learner ? 3

4. (*a*) What is an exhaustive rule-sets in Rule based classification ? If the rule-set is not exhaustive, what problem arises ? How is it resolved ? 4

(*b*) What is progressive sampling ? What are its advantages ? 3

(*c*) State Bayes' theorem. What assumption is used by the Naïve Bayes classifier ? 3

5. (*a*) Consider the following set of frequent 3-itemsets :

{1, 2, 3}, {1, 2, 4}, {1, 2, 5}{1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}.

Assume that there are only five items in the dataset.

(*i*) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

(*ii*) List all candidate 4-itemsets obtained by a candidate generation procedure in Apriori. 6

(*b*) Let X denotes the categorical attribute having values {awful, poor, OK, good}. What is the representation of each value when X is converted to binary form using :

(*i*) 2 bits

(*ii*) 4 bits ? 4

6. Consider the following transactional dataset : 8

Transaction ID	Items Bought
0001	{*a*, *d*, *e*}
0002	{*a*, *b*, *c*, *e*}
0003	{*a*, *b*, *d*, *e*}
0004	{*a*, *c*, *d*, *e*}
0005	{*b*, *c*, *e*}
0006	{*b*, *d*, *e*}
0007	{*c*, *d*}
0008	{*a*, *b*, *c*}
0009	{*a*, *d*, *e*}
0010	{*a*, *b*, *e*}

(*i*) Find out the support of itemsets $\{e\}$; $\{b, d\}$, $\{a, d\}$ and $\{b, d, e\}$. Are these itemsets frequent if minimum support threshold is 30% ?

(*ii*) Find all the rules generated from the 3-itemset $\{b, d, e\}$. List down the strong rules among these rules if minimum confidence threshold is 60%.

(*b*) What is the difference between nominal attributes and ordinal attributes ? Give an example of each. 2

7. (*a*) Explain the following terms with reference to the DBSCAN clustering algorithm :

(*i*) Core point

(*ii*) Noise point

(*iii*) Border point 6

(*b*) Given the following data points : 2, 4, 10, 12, 3, 20, 30, 11, 25. Assume K = 3 and initial means 2, 4, 6. Show the clusters obtained using K-means algorithm after two iterations and show the new means for the next iteration. 4

[This question paper contains 8 printed pages.]

Your Roll No. 20035570003

Sr. No. of Question Paper : 4711 E

Unique Paper Code : 32347611

Name of the Paper : Data Mining

Name of the Course : **B.Sc. (Hons.) Computer Science**

Semester : VI

Duration : 3 Hours Maximum Marks : 75

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. Question No. **1 (Section A)** is compulsory.
3. Attempt any **4** Questions from Nos. **2** to **8 (Section B)**.
4. Parts of a question must be answered together.

Section A

1. (a) Determine the attribute type for the following : (2)

(i) Bronze, Silver, Gold medals awarded at Olympics

(ii) Number of patients in hospital

(iii) Car color

(iv) Dates in a Calendar

(b) List two applications where graph data structure is used to model the data. (2)

(c) Consider an association rule between items from market basket domain which has high support and high confidence. What does it signify? (2)

(d) Explain the following terms with respect to a density-based clustering algorithm: Core point and Border point (2)

(e) Consider a categorical attribute with five values {awful, poor, OK, good, great}. Covert this attribute to asymmetric binary attributes. (3)

(f) List any two ways in which a noise object differs from an outlier? Explain with the help of the example. (4)

(g) Consider the given dataset with two attributes Age and Salary measured on different scales. What problems might arise if the dataset is directly used for k-means clustering? What steps will you suggest to handle the problem? (4)

	Age (in years)	Salary (in rupees)
1	44	72000
2	27	48000
3	30	54000
4	38	61000
5	50	83000
6	37	67000

(h) How is an eager learner classifier different from a lazy learner classifier? Support your answer with an example from both category of classifiers. (4)

(i) Specify whether each of the following activities should fall under the purview of a data mining task or a database query. Justify your answer.

(i) Dividing the customers of a company according to their gender.

(ii) Predicting the future stock price of a company using historical records. (4)

(j) Explain the concept of following types of clustering schemes :

(i) Fuzzy clustering

(ii) Hierarchical based clustering (4)

(k) Consider the following values for two attributes corresponding to four data points: P1(0,2), P2(2,0), P3(3,1), and P4(5,1). Compute the proximity matrix using the metric as Euclidean Distance. (4)

Section B

2. (a) Consider the following dataset for binary classification problem : (6)

Instance	A	B	C	Target Class
1	T	T	1	+
2	T	T	6	+
3	T	F	5	-
4	F	F	4	+
5	F	T	7	-
6	F	T	3	-
7	F	F	8	-
8	T	F	7	+
9	F	T	5	-

Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

(b) For evaluating the performance of a classifier, how does holdout method ~~and 4~~ differ from k-fold cross validation? For k=5 and datapoints- D1, D2, D3, D4, D5, D6, D7, D8, D9, and D10 in the dataset, mention one possible dataset distribution between training and test partition for k-fold cross-validation. (4)

3. Consider the dataset shown below : (10)

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

(i) Estimate the conditional probabilities for *P(Outlook|Yes)*, *P(Temperature|Yes)*, *P(Humidity|Yes)*, *P(Windy|Yes)*, *P(Outlook|No)*, *P(Temperature|No)*, *P(Humidity|No)*, and *P(Windy|No)*.

(ii) Use these estimate of conditional probabilities to predict the class label (*Play Golf*) for a test sample (*Outlook = Rainy, Temperature = Cool, Humidity = High, Windy = True*) using the naive Bayes approach.

4. The DM Pizza Parlour sells pizzas with optional toppings: pepperoni, pineapple, and pickled-onion. Suppose, you have tried five pizzas (P1 to P5) and kept a record of which you liked :

	Pepperoni	pineapple	pickled-Onion	liked
P1	True	True	True	False
P2	True	False	False	True
P3	False	True	True	False
P4	False	True	False	True
P5	True	False	False	True

Show binarization of the above data and use it to calculate Euclidean distances, to demonstrate how the k-Nearest-Neighbor (k-NN) classifier with majority voting would classify a tuple <*False, True, True*>, for k = 1 and k = 3 respectively. (10)

5. Suppose we have height, weight and T-shirt size of some customers and we need to predict the T-shirt size of a new customer given only height and weight information we have. Use k-Nearest-Neighbor classifier to classify the tuple <161, 64>. Assume

k = 5. Note that the data set should be scaled to range [0-1] prior to classification, using min-max normalization. (10)

Height (in cms)	Weight (in kgs)	T-shirt Size
157	58	S
158	59	S
158	63	M
160	59	M
157	56	S
163	60	M
163	61	M
160	64	M
168	64	L
165	61	L
171	62	L
169	63	L

6. Considerthe following data set with nine transactions. Use Apriori algorithm to compute all frequent itemsets of size one and two, considering 1/3 as the minimum support. Also, generate strong association rules using frequent 2-itemsets, considering 0.65 as the minimum confidence. (10)

TID	Items
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

7. Consider the following data set: {4, 8, 12, 20, 32, 36, 48}. Assuming that k = 2, and initial cluster centers for k-means clustering are 32 and 48. Perform the k-means clustering to arrive at final set of cluster solutions. Also, at the end of every iteration, compute the SSE. (10)

8. Use the distance matrix given below to perform hierarchical clustering using single link and show the dendrogram. (10)

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.24	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0

Unique Paper Code : **32347611**
Name of the Paper : **Data Mining**
Name of the Course : **B.Sc. (H) Computer Science**
Semester : **Semester -VI**
Duration of Examination : **Two Hours**
Maximum Marks : **75 Marks**
Year of Admission : **2015, 2016, 2017**

Instructions for Candidates

Attempt Any Four questions. All Questions carry equal marks.

Q 1. Given the following table, classify all the attributes appearing in the table as *binary, discrete or continuous*. Also classify them as qualitative (*nominal or ordinal*) or quantitative (r*atio or interval*). Justify your answer in each case. Show the normalization (scaling data between 0 and 1) of values in the age attribute column. How can you handle missing values in Age column and Height column? Replace Y and N respectively by 1 and 0 in the first column. Replace M and F respectively by 0 and 1 in the second column. You will get one binay vector each for Smoker attribute and Gender attribute. Find out the similarity measure between these two vectors using Jaccard coefficient.

Smoker	Gender	Age	Height	Marital Status
Y	F	32	Tall	Married
Y	M	34	Medium	Marries
N	F	39	Medium	Single
Y	M	41	Tall	Single
Y	M	25	Tall	Divorcee
N	M	36	Tall	Single
Y	F	45	Short	Married
Y	M	31	Tall	Single
N	M	29	Medium	Divorcee
N	F	51	Tall	Single
Y	F	38	Short	Married

Q 2. Given the following binary classification problem :

Instance	A1	A2	Target Class
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-
7	F	F	-
8	T	F	+
9	F	T	-
10	T	F	-

Calculate separately the information gain when splitting is done on A1 and on A2. Which attribute would the decision tree induction algorithm choose? Calculate separately the gain in the Gini index when splitting is done on A1 and A2. Which attribute would the decision tree induction algorithm choose? Is it possible that information gain and the gain in Gini index favour different attributes? Explain your answer.

Q 3. Consider the one dimensional labeled data set given below:

X:	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
Y:	-	-	+	+	-	-	+	+	-	-

Classify the data point x= 5.0 according to its 3- and 5- nearest neighbour using majority vote.

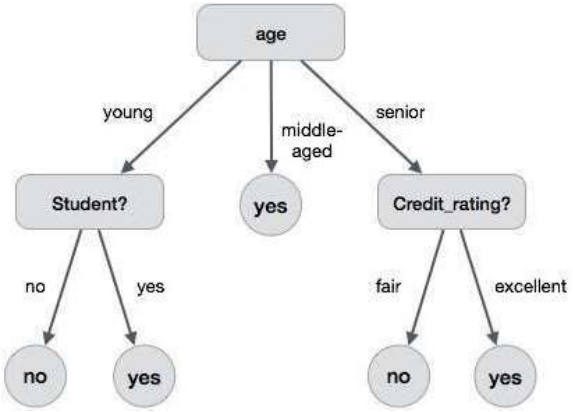
Suppose that there are a total of 60 data mining related documents in a library of 200 documents. Suppose that a search engine retrieves 20 documents after a user enters ‘data mining’ as a query, of which 5 are data mining related documents. What are the precision and recall?

Q 4. Consider the Market Basket dataset shown below:

Transaction ID	Items Bought
0001	{a,d,e}
0024	{a,b,c,e}
0012	{a,b,d,e}
0031	{a,c,d,e}
0015	{b,c,e}
0022	{b,d,e}
0029	{c,d}
0040	{a,b,c}
0033	{a,d,e}
0038	{a,b,e}

Compute support for item sets {e}, {b,d}, {b,d,e}, {a,b,c,e} and {a,b,c,d,e}. Find all association rules that can be generated from item set {b,d,e} along with the confidence of each rule. Is confidence a symmetric measure?

Q 5. Generate all the rules given the following decision tree:



Arrange the generated rule according to class based ordering and classify the following tuples:

Age = young	student = no	loan_approval=?
Age = senior	credit_rating = excellent	loan_approval=?

Consider a training set that contains 90 positive examples and 300 negative examples. For each of the following rules:

R1: A----->+ (Covers 30 positive and 10 negative examples)

R2:B---->+ (Covers 90 positive and 80 negative examples)

R3:C ---->+ (Covers 4 positive and 1 negative example)

Determine which is the best and which is the worst candidate rule according to rule Accuracy.

Q 6. Use K-means algorithm and Euclidean distance to cluster five data points (A4-A8) given below, into 3 clusters. The coordinates of the data points are:

A1(2,8), A2(2,5), A3(1,2), A4(5,8), A5(7,3), A6(6,4), A7(8,4), A8(4,7).

Use A1, A2, A3 as initial centroids. For which situations K-mean clustering will give good results and when will it fail to produce good results?

[This question paper contains 7 printed pag

Your Roll

Sr. No. of Question Paper : 1174 **A**

Unique Paper Code : 32347611

Name of the Paper : Data Mining

Name of the Course : **B.Sc. (Hons.) Computer Science**

Semester : VI

Duration : 3 Hours Maximum Marks : 75

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. Question No. **1 (Section A)** is compulsory.
3. Attempt any **4** Questions from Nos. **2** to **8 (Section B)**.
4. Parts of a question must be answered together.

Section A

1. (a) How are accuracy rate and error calculated for evaluation of a classification model? (2)

P.T.O.

(b) Briefly describe the aggregation technique in data-preprocessing? (2)

(c) Normalize the age of four students, given by the values {18, 21, 22, 25}. (2)

(d) Explain briefly the significance of dimensionality reduction. (2)

(e) What is an outlier in context of a dataset? (2)

(f) What kind of Association Rules do you think would be stronger and more interesting – the rules with high support and low confidence or the rules with low support and high confidence? Why? (3)

(g) Define the use of sampling in data mining? Name two sampling methods. (3)

(h) What are the three factors that affect the computational complexity of Apriori algorithm? (3)

(i) Distinguish between the following type of clustering schemes :

(i) Exclusive vs. Fuzzy Clustering

(ii) Complete vs. Partial Clustering (4)

(j) What do you understand by the term missing data in data mining? Briefly describe two methods for dealing with missing data. (4)

Define the terms scalability and heterogeneity? What challenges do they pose while mining the data? (4)

(l) Define precision and recall metrics used for classification. (4)

Section B

2. (a) Explain discretization and binarization in context of data pre-processing. (4)

(b) Consider a categorical attribute Customer satisfaction {unsatisfactory, poor, neutral, good, very good}

(i) Convert the above categorical attribute to three binary attributes. (2)

(ii) Convert the same attribute to five asymmetric binary attributes. (2)

(c) State the Apriori Principle. (2)

3. For the given employee table, identify the type of each attribute (nominal, ordinal, interval- scaled, ratio-scaled), giving justification for your choice. For each attribute that has missing values, briefly state how will you handle missing values therein. (10)

Emp_id	Gender	Age	Home _pin _code	Date_ of_ joining	Desig.	Contact_No	Email_id
1001	M	32	232322	16/4/10	Captain	981828706	b@gma.com
1002	F	31	222321	21/3/11	Captain	981121072	f@gma.com
1003	F	34	243431	23/4/08	Major	992665007	??
1004	M	??	232432	21/5/09	Captain	987654390	r@gma.com
1005	M	35	454656	13/4/07	Colonel	981123456	d@gma.com
1006	??	36	465645	04/5/05	Colonel	786789564	a@gma.com
1007	F	30	234123	09/7/12	Captain	885678909	??
1008	M	32	676878	18/7/10	Major	??	x@gma.com
1009	M	33	565768	24/6/11	Colonel	989967890	e@gma.com
1010	M	30	498976	05/9/12	Major	??	d@gma.com

4. (a) Consider the following dataset where each data object has a class label along with five features associated with it.

Class	Cap Shape	Bruises	Odour	Stalk Shape	Habitat
Edible	Flat	Yes	anise	Tapering	grasses
poisonous	Convex	Yes	pungent	enlargening	grasses
Edible	Convex	Yes	almond	enlargening	grasses
Edible	Convex	Yes	almond	Tapering	meadows
Edible	Flat	Yes	anise	enlargening	woods
Edible	flat	No	none	enlargening	urban
poisonous	conical	Yes	pungent	enlargening	urban
Edible	flat	Yes	anise	enlargening	meadows
poisonous	convex	Yes	pungent	enlargening	urban

Consider the following pair of rules :

- *(Odour = pungent) and (habitat = urban)* → (Class = *poisonous)*
- *(Bruises = yes)* → *(Class = edible)*

(i) Are the two rules mutually exclusive? Justify your answer. (2)

(ii) Calculate coverage and accuracy for each of the rules. (4)

(b) Consider the one-dimensional labeled data set given below :

X:	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
Y:	-	-	+	+	-	-	+	+	-	-

Classify the data point x = 4.0 according to the 5-nearest neighbours, using the majority voting scheme. (4)

5. (a) What are the three conditions needed to be satisfied by a distance measure, so that it can be established as a distance metric? (3)

(b) Show whether Euclidean Distance, used for finding distance between two data objects $o_1(x_1, y_1)$ and $o_2(x_2, y_2)$, can be treated as a distance metric. (6)

(c) With the help of a diagram, explain the usage of a dendrogram. (1)

6. Consider a transaction database D, consisting of nine transactions, as shown in the following table. Suppose the minimum support is set at 45% and the minimum confidence is set at 70%, show clearly the steps for finding out frequent itemsets of all sizes using the Apriori algorithm. Also generate the strong association rules from the frequent itemsets of size 3. (10)

TID	List of Items
T1	A,B,C,F
T2	B,D
T3	B,C
T4	A,B,C
T5	A,C,F
T6	B,C,F
T7	A,D
T8	A,B,C,E,F
T9	A,B,C

7. Consider a dataset of images of dogs and cats. Suppose there are 500 images of dogs and cats each. The classification model predicts 340 correct images of dogs and 410 correct images of cat. Perform the operations that follow :

(a) Draw the confusion matrix for this problem.

(b) Compute the classifier accuracy, error and sensitivity. (4+6)

8. Given the following data points: 4, 9, 18, 13, 11, 2, 6, 25, k = 3 and initial centroids $\mu_1 = 5$, $\mu_2 = 10$ and $\mu_3 = 15$. Show clearly the clusters and new cluster centres obtained after each iteration of K-means algorithm for two iterations of the algorithm. (10)

This question paper contains 8 printed pages]

Roll No. | | | | | | | | | | |

S. No. of Question Paper : 9426A

Unique Paper Code : 32347611 HC

Name of the Paper : Data Mining

Name of the Course : B.Sc. (H) Computer Science : DSE-4

Semester : VI

Duration : 3 Hours Maximum Marks : 75

(Write your Roll No. on the top immediately on receipt of this question paper.)

All questions are compulsory from Section A.

Attempt any *four* questions from Section B.

Section A

1. (*a*) What is the difference between Data mining and KDD ? 2

(*b*) Identify attribute types for the following : 2

(*i*) eye color

(*ii*) grades

(*iii*) dates in a calendar

(*iv*) age.

P.T.O.

(c) What are the maximum and minimum values of Gini Index? Find Gini index for the following node : 2

Node N	Count
Class = 0	1
Class : 1	5

(d) Give *two* applications where graph data structure is used to model data. 2

(e) Given four points $p_1(0,2)$, $p_2(2,0)$, $p_3(3,1)$ and $p_4(5,1)$. Calculate Euclidian distance between the points p_1 and p_2, and p_3 and p_4. 2

(f) Let X denote the categorical attribute having possible values {poor, good, better, best}. What is the representation of each value when X is converted to binary form ? 2

(g) How are interval scaled attributes different from ratio scaled attributes ? Give an example of each. 3

(*h*) How is a eager learner different from lazy learner ? Support your answer with an example from both categories of classifiers. 3

(*i*) State the *Apriori principle*. Comment on the following statement :

"If an item set {*x*, *y*, *z*} is frequent, then its subset {*y*, *z*} will be frequent." 4

(*j*) Given the age of four students, normalize the values {18, 21, 22, 25}. 4

(*k*) Explain the following terms with reference to the DBSCAN algorithm : 2+2

(*i*) Core point

(*ii*) Noise point

(*l*) What are *mutually exclusive* rules in a rule based classifier ? What problem may arise if rules are not mutually exclusive ? How can such problem be resolved ? 5

P.T.O.

Section B

2. (*a*) Consider the following transaction dataset : 6

Customer ID	Transaction ID	Items Bought
1	0001	{*a*, *d*, *e*}
1	0024	{*a*, *b*, *c*, *e*}
2	0012	{*a*, *b*,*d*,*e*}
2	0031	{*a*, *c*, *d*, *e*}
3	0015	{*b*, *c*, *e*}
3	0022	{*b*, *d*, *e*}
4	0029	{*c*, *d*}
4	0040	{*a*, *b*, *c*}
5	0033	{*a*, *d*, *e*}
5	0038	{*a*, *b*, *e*}

(*i*) Compute the support of itemsets {*e*}, {*b*, *d*}, {*b*, *d*, *e*}, {*a*, *b*, *d*, *e*}

(*ii*) Compute the confidence of rules {*b*, *d*} → {*e*} and {*e*} → {*b*, *d*}.

(*iii*) Is confidence a symmetric measure ?

(*b*) Let A and B be two sets of integers. A distance measure '*d*' is defined as $d(A - B) = \text{size}(A - B)$, where '–' denotes the set difference. Prove that '*d*' is not a metric. 4

3. (*a*) Explain the concept of aggregation with the help of an example. List three uses of aggregation. 2+3

(*b*) What is the difference between noise and outliers ? Answer the following questions : 5

(*i*) Is noise ever interesting or desirable ?

(*ii*) Are outliers ever interesting or desirable ?

(*iii*) Are noise obejcts always outliers ?

(*iv*) Are outliers always noise objects ?

4. (*a*) For the following two-class problem, draw a *Confusion Matrix* and compute the *Accuracy* and *Error* from it : 6

Instance id	A	B	Predicted Class	Actual Class
1	T	F	+	+
2	T	T	+	+
3	T	T	+	–
4	T	F	–	–
5	T	T	+	+
6	F	F	–	+
7	F	F	–	–
8	F	F	–	–
9	T	T	–	+
10	T	F	–	+

P.T.O.

(*b*) What is k - fold cross validation ? How is it different from the hold-out method ? 4

5. (*a*) What is the difference between hierarchical and partition based clustering ? Enumerate *two* advantages and disadvantages of hierarchical clustering. 4

(*b*) What is simple random sampling ? 2

(*c*) Consider the following rule set : 4

R_1 : (Give Birth = no) ^ (Can Fly = yes) → Birds

R_2 : (Give Birth = no) ^ (Live in Water = yes) → Fishes

R_3 : (Give Birth = yes) ^ (Blood Type = warm) → Mammals

R_4 : (Give Birth = no) ^ (Can Fly = no) → Reptiles

R_5 : (Live in Water = sometimes) → Amphibians.

Which rules cover the following tuples :

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	Warm	No	Yes	No	?
bear	Warm	Yes	No	No	?

6. (*a*) List the rules that can be generated from tlre 3- itemset ABE using the following transactional data set. Compute the confidence. 8

T_id	Itemset
t1	ACD
t2	BCE
t3	ABCE
t4	BDE
t5	ABCE
t6	ABCD

(*b*) Enumerate strong association rules if *minConf* = 0.5 2

7. (*a*) Given the following points : 2, 4, 10, 12, 3, 20, 30, 11, 25. Given $k = 3$, and the initial means, $\mu_1 = 2$, $\mu_2 = 4$ and $\mu_3 = 6$. Show the clusters obtained and the new means after each iteration using the K-means algorithm. 8

(*b*) What is the differences between Partial and Complete clustering scheme ? 2